Greedy-Like Algorithms for the Cosparse Analysis Model

R. Giryes[a,*], S. Nam[b], M. Elad[a], R.Gribonval[b], M. E. Davies[c]

[a]*The Department of Computer Science, Technion – Israel Institute of Technology, Haifa 32000, Israel*
[b]*INRIA Rennes - Bretagne Atlantique, Campus de Beaulieu, F-35042 Rennes Cedex, France*
[c]*School of Engineering and Electronics, The University of Edinburgh, The King's Buildings, Mayfield Road, Edinburgh EH9 3JL, UK*

Abstract

The cosparse analysis model has been introduced recently as an interesting alternative to the standard sparse synthesis approach. A prominent question brought up by this new construction is the analysis pursuit problem – the need to find a signal belonging to this model, given a set of corrupted measurements of it. Several pursuit methods have already been proposed based on ℓ_1 relaxation and a greedy approach. In this work we pursue this question further, and propose a new family of pursuit algorithms for the cosparse analysis model, mimicking the greedy-like methods – compressive sampling matching pursuit (CoSaMP), subspace pursuit (SP), iterative hard thresholding (IHT) and hard thresholding pursuit (HTP). Assuming the availability of a near optimal projection scheme that finds the nearest cosparse subspace to any vector, we provide performance guarantees for these algorithms. Our theoretical study relies on a restricted isometry property adapted to the context of the cosparse analysis model. We explore empirically the performance of these algorithms by adopting a plain thresholding projection, demonstrating their good performance.

Keywords: Sparse representations, Compressed sensing, Synthesis, Analysis, CoSaMP, Subspace-pursuit, Iterative hard threshodling, Hard thresholding pursuit.
2010 MSC: 94A20, 94A12, 62H12

1. Introduction

Many natural signals and images have been observed to be inherently low dimensional despite their possibly very high ambient signal dimension. It is by now well understood that this phenomenon lies at the heart of the success of numerous methods of signal and image processing. Sparsity-based models for signals offer an elegant and clear way to enforce such inherent low-dimensionality, explaining their high popularity in recent years. These models consider the signal $\mathbf{x} \in \mathbb{R}^d$ as belonging to a finite union of subspaces of dimension $k \ll d$ [1]. In this paper we shall focus on one such approach – the cosparse analysis model – and develop pursuit methods for it.

Before we dive into the details of the model assumed and the pursuit problem, let us first define the following generic inverse problem that will accompany us throughout the paper: For some unknown signal $\mathbf{x} \in \mathbb{R}^d$, an incomplete set of linear observations $\mathbf{y} \in \mathbb{R}^m$ (incomplete implies $m < d$) is available via

$$\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{e}, \tag{1}$$

where $\mathbf{e} \in \mathbb{R}^m$ is an additive bounded noise that satisfies $\|\mathbf{e}\|_2^2 \leq \epsilon^2$. The task is to recover or approximate $\mathbf{x}$. In the noiseless setting where $\mathbf{e} = 0$, this amounts to solving $\mathbf{y} = \mathbf{M}\mathbf{x}$. Of course, a simple fact in linear algebra tells us that this problem admits infinitely many solutions (since $m < d$). Therefore, when all we have is the observation $\mathbf{y}$ and the measurement/observation matrix $\mathbf{M} \in \mathbb{R}^{m \times d}$, we are in a hopeless situation to recover $\mathbf{x}$.

*Corresponding author

1.1. The Synthesis Approach

This is where 'sparse signal models' come into play. In the sparse synthesis model, the signal $\mathbf{x}$ is assumed to have a very sparse representation in a given fixed dictionary $\mathbf{D} \in \mathbb{R}^{d\times n}$. In other words, there exists $\boldsymbol{\alpha}$ with few nonzero entries, as counted by the "ℓ_0-norm" $\|\boldsymbol{\alpha}\|_0$, such that

$$\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}, \quad \text{and} \quad k := \|\boldsymbol{\alpha}\|_0 \ll d. \tag{2}$$

Having this knowledge we solve (1) using

$$\hat{\mathbf{x}}_{\ell_0} = \mathbf{D}\hat{\boldsymbol{\alpha}}_{\ell_0}, \quad \text{and} \quad \hat{\boldsymbol{\alpha}}_{\ell_0} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\boldsymbol{\alpha}\|_0 \quad \text{subject to} \quad \|\mathbf{y} - \mathbf{MD}\boldsymbol{\alpha}\|_2 \le \epsilon. \tag{3}$$

More details about the properties of this problem can be found in [2, 3].

Since solving (3) is an NP-complete problem [4], approximation techniques are required for recovering $\mathbf{x}$. One strategy is by using relaxation, replacing the ℓ_0 with ℓ_1 norm, resulting with the ℓ_1-synthesis problem

$$\hat{\mathbf{x}}_{\ell_1} = \mathbf{D}\hat{\boldsymbol{\alpha}}_{\ell_1}, \quad \text{and} \quad \hat{\boldsymbol{\alpha}}_{\ell_1} = \operatorname{argmin} \|\boldsymbol{\alpha}\|_1 \quad s.t. \quad \|\mathbf{y} - \mathbf{MD}\boldsymbol{\alpha}\|_2 \le \epsilon. \tag{4}$$

For a unitary matrix $\mathbf{D}$ and a vector $\mathbf{x}$ with k-sparse representation $\boldsymbol{\alpha}$, if $\delta_{2k} < \delta_{\ell_1}$ then

$$\left\|\hat{\mathbf{x}}_{\ell_1} - \mathbf{x}\right\|_2 \le C_{\ell_1} \|\mathbf{e}\|_2 , \tag{5}$$

where $\hat{\mathbf{x}}_{\ell_1} = \mathbf{D}\hat{\boldsymbol{\alpha}}_{\ell_1}$, δ_{2k} is the constant of the restricted isometry property (RIP) of $\mathbf{MD}$ for 2k sparse signals, C_{ℓ_1} is a constant greater than $\sqrt{2}$ and δ_{ℓ_1} ($\simeq 0.4931$) is a reference constant [5, 6, 7]. Note that this result implies a perfect recovery in the absence of noise. The above statement was extended also for incoherent redundant dictionaries [8].

Another option for approximating (3) is using a greedy strategy, like in the thresholding technique or orthogonal matching pursuit (OMP) [9, 10]. A different related approach is the greedy-like family of algorithms. Among those we have compressive sampling matching pursuit (CoSaMP) [11], subspace pursuit (SP) [12], iterative hard thresholding (IHT) [13] and hard thresholding pursuit (HTP) [14]. CoSaMP and SP were the first greedy methods shown to have guarantees in the form of (5) assuming $\delta_{4k} < \delta_{\text{CoSaMP}}$ and $\delta_{3k} \le \delta_{\text{SP}}$ [11, 12, 6, 15]. Following their work, iterative hard thresholding (IHT) and hard thresholding pursuit (HTP) were shown to have similar guarantees under similar conditions [13, 14, 16, 6]. Recently, a RIP based guarantee was developed also for OMP [17].

1.2. The Cosparse Analysis Model

Recently, a new signal model called *cosparse analysis model* was proposed in [18, 19]. The model can be summarized as follows: For a fixed analysis operator $\boldsymbol{\Omega} \in \mathbb{R}^{p\times d}$ referred to as the analysis dictionary, a signal $\mathbf{x} \in \mathbb{R}^d$ belongs to the cosparse analysis model with cosparsity ℓ if

$$\ell := p - \|\boldsymbol{\Omega}\mathbf{x}\|_0 . \tag{6}$$

The quantity ℓ is the number of rows in $\boldsymbol{\Omega}$ that are orthogonal to the signal. The signal $\mathbf{x}$ is said to be ℓ-cosparse, or simply cosparse. We denote the indices of the zeros of the analysis representation as the *cosupport* Λ and the submatrix that contains the rows from $\boldsymbol{\Omega}$ that belong to Λ by $\boldsymbol{\Omega}_\Lambda$. As the definition of cosparsity suggests, the emphasis of the cosparse analysis model is on the zeros of the analysis representation vector $\boldsymbol{\Omega}\mathbf{x}$. This contrasts the emphasis on 'few non-zeros' in the synthesis model (2). It is clear that in the case where every ℓ rows in $\boldsymbol{\Omega}$ are independent, $\mathbf{x}$ resides in a subspace of dimension $d - \ell$ that consists of vectors orthogonal to the rows of $\boldsymbol{\Omega}_\Lambda$. In the general case where dependencies occur between the rows of $\boldsymbol{\Omega}$, the dimension is d minus the rank of $\boldsymbol{\Omega}_\Lambda$. This is similar to the behavior in the synthesis case where a k-sparse signal lives in a k-dimensional space. Thus, for this model to be effective, we assume a large value of ℓ.

In the analysis model, recovering $\mathbf{x}$ from the corrupted measurements is done by solving the following minimization problem [20]:

$$\mathbf{x}_{A-\ell_0} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\boldsymbol{\Omega}\mathbf{x}\|_0 \quad \text{subject to} \quad \|\mathbf{y} - \mathbf{M}\mathbf{x}\|_2 \le \epsilon. \tag{7}$$

Solving this problem is NP-complete [18], just as in the synthesis case, and thus approximation methods are required. As before, we can use an ℓ_1 relaxation to (7), replacing the ℓ_0 with ℓ_1 in (7), resulting with the ℓ_1-analysis problem

[18, 20, 21, 22]. Another option is the greedy approach. A greedy algorithm called Greedy Analysis Pursuit (GAP) has been developed in [18, 19, 23] that somehow mimics Orthogonal Matching Pursuit [9, 10] with a form of iterative reweighted least Squares (IRLS) [24]. Other alternatives for OMP, backward greedy (BG) and orthogonal BG (OBG), were presented in [25] for the case that $\mathbf{M}$ is the identity. For the same case, the parallel to the thresholding technique was analyzed in [26].

1.3. This Work

Another avenue exists for the development of analysis pursuit algorithms – constructing methods that will imitate the family of greedy-like algorithms. Indeed, we have recently presented preliminary and simplified versions of analysis IHT (AIHT), analysis HTP (AHTP), analysis CoSaMP (ACoSaMP) and Analysis SP (ASP) in [27, 28] as analysis versions of the synthesis counterpart methods. This paper re-introduces these algorithms in a more general form, ties them to their synthesis origins, and analyze their expected performance. The main contribution of the paper is our result on the stability of these analysis pursuit algorithms. We show that after a finite number of iterations and for a given constant c_0, the reconstruction result $\hat{\mathbf{x}}$ of AIHT, AHTP, ACoSaMP and ASP all satisfy

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq c_0 \|\mathbf{e}\|_2 \,, \tag{8}$$

under a RIP-like condition on $\mathbf{M}$ and the assumption that we are given a good near optimal projection scheme. A bound is also given for the case where $\mathbf{x}$ is only nearly ℓ-cosparse. Similar results for the ℓ_1 analysis appear in [21, 22]. More details about the relation between these papers and our results will be given in Section 6. In addition to our theoretical results we demonstrate the performance of the four pursuit methods under a thresholding based simple projection scheme. Both our theoretical and empirical results show that linear dependencies in $\mathbf{\Omega}$ that result with a larger cosparsity in the signal $\mathbf{x}$, lead to a better reconstruction performance. *This suggests that, as opposed to the synthesis case, strong linear dependencies within $\mathbf{\Omega}$ are desired.*

This paper is organized as follows:

- In Section 2 we present the notation used along the paper.
- In Section 3 we define a RIP-like property, the $\mathbf{\Omega}$-RIP, for the analysis model, proving that it has similar characteristics like the regular RIP. In Section 4 the notion of near optimal projection is proposed and some nontrivial operators for which a tractable optimal projection exists are exhibited. Both the $\mathbf{\Omega}$-RIP and the near optimal projection are used throughout this paper as a main force for deriving our theoretical results.
- In Section 5 the four pursuit algorithms for the cosparse analysis framework are defined, adapted to the general format of the pursuit problem we have defined above.
- In Section 6 we derive the success guarantees for all the above algorithms in a unified way. Note that the provided results can be easily adapted to other union-of-subspaces models given near optimal projection schemes for them, in the same fashion done for IHT with an optimal projection scheme in [29]. The relation between the obtained results and existing work appears in this section as well.
- Empirical performance of these algorithms is demonstrated in Section 7 in the context of the cosparse signal recovery problem. We use a simple thresholding as the near optimal projection scheme in the greedy-like techniques.
- Section 8 discuss the presented results and concludes our work.

2. Notations and Preliminaries

We use the following notation in our work:

- $\sigma_{\mathbf{M}}$ is the largest singular value of $\mathbf{M}$, i.e., $\sigma_{\mathbf{M}}^2 = \|\mathbf{M}^*\mathbf{M}\|_2$.
- $\|\cdot\|_2$ is the euclidian norm for vectors and the spectral norm for matrices. $\|\cdot\|_1$ is the ℓ_1 norm that sums the absolute values of a vector and $\|\cdot\|_0$, though not really a norm, is the ℓ_0-norm which counts the number of non-zero elements in a vector.

- Given a cosupport set Λ, $\mathbf{\Omega}_\Lambda$ is a sub-matrix of $\mathbf{\Omega}$ with the *rows* that belong to Λ.
- For given vectors $\mathbf{v}, \mathbf{z} \in \mathbb{R}^d$ and an analysis dictionary $\mathbf{\Omega}$, cosupp($\mathbf{\Omega v}$) returns the cosupport of $\mathbf{\Omega v}$ and cosupp($\mathbf{\Omega z}, \ell$) returns the index set of ℓ smallest (in absolute value) elements in $\mathbf{\Omega z}$. If more than ℓ elements are zero all of them are returned. In the case where the ℓ-th smallest entry is equal to the $\ell + 1$ smallest entry, one of them is chosen arbitrarily.
- In a similar way, in the synthesis case $\mathbf{D}_T$ is a sub-matrix of $\mathbf{D}$ with *columns*[1] corresponding to the set of indices T, supp$(\cdot)$ returns the support of a vector, supp$(\cdot, k)$ returns the set of k-largest elements and $\lceil\cdot\rceil_k$ preserves the k-largest elements in a vector. In the case where the k-th largest entry is equal to the $k + 1$ largest entry, one of them is chosen arbitrarily.
- $\mathbf{Q}_\Lambda = \mathbf{I} - \mathbf{\Omega}_\Lambda^\dagger \mathbf{\Omega}_\Lambda$ is the orthogonal projection onto the orthogonal complement of range$(\mathbf{\Omega}_\Lambda^*)$.
- $\mathbf{P}_\Lambda = \mathbf{I} - \mathbf{Q}_\Lambda = \mathbf{\Omega}_\Lambda^\dagger \mathbf{\Omega}_\Lambda$ is the orthogonal projection onto range$(\mathbf{\Omega}_\Lambda^*)$.
- $\hat{\mathbf{x}}_{\text{AIHT}}/\hat{\mathbf{x}}_{\text{AHTP}}/\hat{\mathbf{x}}_{\text{ACoSaMP}}/\hat{\mathbf{x}}_{\text{ASP}}$ are the reconstruction results of AIHT/ AHTP/ ACoSaMP/ ASP respectively. Sometimes when it is clear from the context to which algorithms we refer, we abuse notations and use $\hat{\mathbf{x}}$ to denote the reconstruction result.
- A cosupport Λ has a corank r if rank$(\mathbf{\Omega}_\Lambda) = r$. A vector $\mathbf{v}$ has a corank r if its cosupport has a corank r.
- $[p]$ denotes the set of integers $[1 \dots p]$.
- $\mathcal{L}_{\mathbf{\Omega},\ell} = \{\Lambda \subseteq [p], |\Lambda| \geq \ell\}$ is the set of ℓ-cosparse cosupports and $\mathcal{L}^{\text{corank}}_{\mathbf{\Omega},r} = \{\Lambda \subseteq [p], \text{rank}(\mathbf{\Omega}_\Lambda) \geq r\}$ is the set of all cosupports with corresponding corank r.
- $\mathcal{W}_\Lambda = \text{span}^\perp(\mathbf{\Omega}_\Lambda) = \{\mathbf{Q}_\Lambda \mathbf{z}, \mathbf{z} \in \mathbb{R}^d\}$ is the subspace spanned by a cosparsity set Λ.
- $\mathcal{A}_{\mathbf{\Omega},\ell} = \bigcup_{\Lambda \in \mathcal{L}_{\mathbf{\Omega},\ell}} \mathcal{W}_\Lambda$ is the union of subspaces of ℓ-cosparse vectors and $\mathcal{A}^{\text{corank}}_{\mathbf{\Omega},r} = \bigcup_{\Lambda \in \mathcal{L}^{\text{corank}}_{\mathbf{\Omega},r}} \mathcal{W}_\Lambda$ is the union of subspaces of all vectors with corank r. In the case that every ℓ rows of $\mathbf{\Omega}$ are independent it is clear that $\mathcal{A}_{\mathbf{\Omega},\ell} = \mathcal{A}^{\text{corank}}_{\mathbf{\Omega},r}$. When it will be clear from the context, we will remove $\mathbf{\Omega}$ from the subscript.
- $\mathbf{x} \in \mathbb{R}^d$ denotes the original unknown ℓ-cosparse vector and $\Lambda_{\mathbf{x}}$ its cosupport.
- $\mathbf{v}, \mathbf{u} \in \mathcal{A}_\ell$ are used to denote general ℓ-cosparse vectors and $\mathbf{z} \in \mathbb{R}^d$ is used to denote a general vector.

3. Ω-RIP Definition and its Properties

We now turn to define the $\mathbf{\Omega}$-RIP, which parallels the regular RIP as used in [5]. This property is a very important property for the analysis of the algorithms which holds for a large family of matrices $\mathbf{M}$ as we will see hereafter.

Definition 3.1. *A matrix* $\mathbf{M}$ *has the* $\mathbf{\Omega}$*-RIP property with a constant* δ_ℓ*, if* δ_ℓ *is the smallest constant that satisfies*

$$(1 - \delta_\ell) \|\mathbf{v}\|_2^2 \leq \|\mathbf{M}\mathbf{v}\|_2^2 \leq (1 + \delta_\ell) \|\mathbf{v}\|_2^2 , \tag{9}$$

whenever $\mathbf{\Omega v}$ *has at least* ℓ *zeroes.*

Note that though δ_ℓ is also a function of $\mathbf{\Omega}$ we abuse notation and use the same symbol for the $\mathbf{\Omega}$-RIP as the regular RIP. It will be clear from the context to which of them we refer and what $\mathbf{\Omega}$ is in use with the $\mathbf{\Omega}$-RIP. A similar property that looks at the corank of the vectors can be defined

[1] By the abuse of notation we use the same notation for the selection sub-matrices of rows and columns. The selection will be clear from the context since in analysis the focus is always on the rows and in synthesis on the columns.

Definition 3.2. *A matrix* $\mathbf{M}$ *has the corank-*$\mathbf{\Omega}$*-RIP property with a constant* δ_r^{corank}*, if* δ_r^{corank} *is the smallest constant that satisfies*

$$(1-\delta_r^{corank})\,\|\mathbf{u}\|_2^2 \le \|\mathbf{M}\mathbf{u}\|_2^2 \le (1+\delta_r^{corank})\,\|\mathbf{u}\|_2^2 \tag{10}$$

whenever the corank of $\mathbf{u}$ *with respect to* $\mathbf{\Omega}$ *is greater or equal to* r*.*

The $\mathbf{\Omega}$-RIP, like the regular RIP, inherits several key properties. We present only those related to δ_ℓ, while very similar characteristics can be derived also for the corank-$\mathbf{\Omega}$-RIP. The first property we pose is an immediate corollary of the δ_ℓ definition.

Corollary 3.3. *If* $\mathbf{M}$ *satisfies the* $\mathbf{\Omega}$*-RIP with a constant* δ_ℓ *then*

$$\|\mathbf{M}\mathbf{Q}_\Lambda\|_2^2 \le 1+\delta_\ell \tag{11}$$

for any $\Lambda \in \mathcal{L}_\ell$*.*

Proof: Any $\mathbf{v} \in \mathcal{A}_\ell$ can be represented as $\mathbf{v} = \mathbf{Q}_\Lambda \mathbf{z}$ with $\Lambda \in \mathcal{L}_\ell$ and $\mathbf{z} \in \mathbb{R}^d$. Thus, the $\mathbf{\Omega}$-RIP in (9) can be reformulated as

$$(1-\delta_\ell)\,\|\mathbf{Q}_\Lambda \mathbf{z}\|_2^2 \le \|\mathbf{M}\mathbf{Q}_\Lambda \mathbf{z}\|_2^2 \le (1+\delta_\ell)\,\|\mathbf{Q}_\Lambda \mathbf{z}\|_2^2 \tag{12}$$

for any $\mathbf{z} \in \mathbb{R}^d$ and $\Lambda \in \mathcal{L}_\ell$. Since $\mathbf{Q}_\Lambda$ is a projection $\|\mathbf{Q}_\Lambda \mathbf{z}\|_2^2 \le \|\mathbf{z}\|_2^2$. Combining this with the right inequality in (12) gives

$$\|\mathbf{M}\mathbf{Q}_\Lambda \mathbf{z}\|_2^2 \le (1+\delta_\ell)\,\|\mathbf{z}\|_2^2 \tag{13}$$

for any $\mathbf{z} \in \mathbb{R}^d$ and $\Lambda \in \mathcal{L}_\ell$. The first inequality in (11) follows from (13) by the definition of the spectral norm. □

Lemma 3.4. *For* $\tilde{\ell} \le \ell$ *it holds that* $\delta_\ell \le \delta_{\tilde{\ell}}$*.*

Proof: Since $\mathcal{A}_\ell \subseteq \mathcal{A}_{\tilde{\ell}}$ the claim is immediate. □

Lemma 3.5. $\mathbf{M}$ *satisfies the* $\mathbf{\Omega}$*-RIP if and only if*

$$\|\mathbf{Q}_\Lambda(\mathbf{I}-\mathbf{M}^*\mathbf{M})\mathbf{Q}_\Lambda\|_2 \le \delta_\ell \tag{14}$$

for any $\Lambda \in \mathcal{L}_\ell$*.*

Proof: The proof is similar to the one of the regular RIP as appears in [6]. As a first step we observe that Definition 3.1 is equivalent to requiring

$$\left|\|\mathbf{M}\mathbf{v}\|_2^2 - \|\mathbf{v}\|_2^2\right| \le \delta_\ell\,\|\mathbf{v}\|_2^2 \tag{15}$$

for any $\mathbf{v} \in \mathcal{A}_\ell$. The latter is equivalent to

$$\left|\|\mathbf{M}\mathbf{Q}_\Lambda \mathbf{z}\|_2^2 - \|\mathbf{Q}_\Lambda \mathbf{z}\|_2^2\right| \le \delta_\ell\,\|\mathbf{Q}_\Lambda \mathbf{z}\|_2^2 \tag{16}$$

for any set $\Lambda \in \mathcal{L}_\ell$ and any $\mathbf{z} \in \mathbb{R}^d$, since $\mathbf{Q}_\Lambda \mathbf{z} \in \mathcal{A}_\ell$. Next we notice that

$$\|\mathbf{M}\mathbf{Q}_\Lambda \mathbf{z}\|_2^2 - \|\mathbf{Q}_\Lambda \mathbf{z}\|_2^2 = \mathbf{z}^*\mathbf{Q}_\Lambda\mathbf{M}^*\mathbf{M}\mathbf{Q}_\Lambda\mathbf{z} - \mathbf{z}^*\mathbf{Q}_\Lambda\mathbf{z} = \langle \mathbf{Q}_\Lambda(\mathbf{M}^*\mathbf{M}-\mathbf{I})\mathbf{Q}_\Lambda\mathbf{z}, \mathbf{z}\rangle.$$

Since $\mathbf{Q}_\Lambda(\mathbf{M}^*\mathbf{M}-\mathbf{I})\mathbf{Q}_\Lambda$ is Hermitian we have that

$$\max_{\mathbf{z}} \frac{|\langle \mathbf{Q}_\Lambda(\mathbf{M}^*\mathbf{M}-\mathbf{I})\mathbf{Q}_\Lambda\mathbf{z}, \mathbf{z}\rangle|}{\|\mathbf{z}\|_2} = \|\mathbf{Q}_\Lambda(\mathbf{M}^*\mathbf{M}-\mathbf{I})\mathbf{Q}_\Lambda\|_2\,. \tag{17}$$

Thus we have that Definition 3.1 is equivalent to (14) for any set $\Lambda \in \mathcal{L}_\ell$. □

Corollary 3.6. *If* $\mathbf{M}$ *satisfies the* $\mathbf{\Omega}$*-RIP then*

$$\left\|\mathbf{Q}_{\Lambda_1}(\mathbf{I}-\mathbf{M}^*\mathbf{M})\mathbf{Q}_{\Lambda_2}\right\|_2 \leq \delta_\ell, \tag{18}$$

for any Λ_1 *and* Λ_2 *such that* $\Lambda_1 \cap \Lambda_2 \in \mathcal{L}_\ell$.

Proof: Since $\Lambda_1 \cap \Lambda_2 \subseteq \Lambda_1$ and $\Lambda_1 \cap \Lambda_2 \subseteq \Lambda_2$

$$\left\|\mathbf{Q}_{\Lambda_1}(\mathbf{I}-\mathbf{M}^*\mathbf{M})\mathbf{Q}_{\Lambda_2}\right\|_2 \leq \left\|\mathbf{Q}_{\Lambda_2\cap\Lambda_1}(\mathbf{I}-\mathbf{M}^*\mathbf{M})\mathbf{Q}_{\Lambda_2\cap\Lambda_1}\right\|_2 .$$

Using Lemma 3.5 completes the proof. □

As we will see later, we require the $\mathbf{\Omega}$-RIP to be small. Thus, we are interested to know for what matrices this hold true. In the synthesis case, where $\mathbf{\Omega}$ is unitary and the $\mathbf{\Omega}$-RIP is identical to the RIP, it was shown for certain family of random matrices, such as matrices with Bernoulli or Subgaussian ensembles, that for any value of ϵ_k if $m \geq C_{\epsilon_k} k \log(\frac{m}{k\epsilon_k})$ then $\delta_k \leq \epsilon_k$ [5, 8, 30], where δ_k is the RIP constant and C_{ϵ_k} is a constant depending on ϵ_k and $\mathbf{M}$. A similar result for the same family of random matrices holds for the analysis case. The result is a special case of the result presented in [29].

Theorem 3.7 (Theorem 3.3 in [29]). *Let* $\mathbf{M} \in \mathbb{R}^{m\times d}$ *be a random matrix such that for any* $\mathbf{z} \in \mathbb{R}^d$ *and* $0 < \tilde{\epsilon} \leq \frac{1}{3}$ *it satisfies*

$$P\left(\left|\|\mathbf{M}\mathbf{z}\|_2^2 - \|\mathbf{z}\|_2^2\right| \geq \tilde{\epsilon}\,\|\mathbf{z}\|_2^2\right) \leq e^{-\frac{C_{\mathbf{M}} m\tilde{\epsilon}}{2}}, \tag{19}$$

where $C_{\mathbf{M}} > 0$ *is a constant. For any value of* $\epsilon_\ell > 0$*, if*

$$m \geq \frac{32}{C_M \epsilon_r^2}\left(\log(\left|\mathcal{L}_r^{corank}\right|) + (d-r)\log(9/\epsilon_r) + t\right), \tag{20}$$

then $\delta_r^{corank} \leq \epsilon_r$ *with probability exceeding* $1 - e^{-t}$.

The above theorem is important since it shows that the $\mathbf{\Omega}$-RIP holds with a small constant for a large family of matrices – the same family that satisfy the RIP property. In a recent work it was even shown that by randomizing the signs of the columns in the matrices that satisfy the RIP we get new matrices that also satisfy the RIP [31]. Thus, requiring the $\mathbf{\Omega}$-RIP constant to be small, as will be done hereafter, is legitimate.

For completeness we present a proof for theorem 3.7 in Appendix A based on [8, 30, 32]. We include in it also the proof of Theorem 3.8 to follow. In the case that $\mathbf{\Omega}$ is in general position $\left|\mathcal{L}_r^{\text{corank}}\right| = \binom{p}{r} \leq (\frac{ep}{p-r})^{p-r}$ (inequality is by Stirling's formula) and thus $m \geq (p-r)\log(\frac{ep}{p-r})$. Since we want m to be smaller than d we need $p - \ell$ to be smaller than d. This limits the size of p for $\mathbf{\Omega}$ since r cannot be greater than d. Thus, we present a variation of the theorem which states the results in terms of δ_ℓ and ℓ instead of δ_r^{corank} and r. The following theorem is also important because of the fact that our theoretical results are in terms of δ_ℓ and not δ_r^{corank}. It shows that δ_ℓ is small in the same family of matrices that guarantees δ_r^{corank} to be small.

Theorem 3.8. *Under the same setup of Theorem 3.7, for any* $\epsilon_\ell > 0$ *if*

$$m \geq \frac{32}{C_M \epsilon_\ell^2}\left((p-\ell)\log\left(\frac{9p}{(p-\ell)\epsilon_\ell}\right) + t\right), \tag{21}$$

then $\delta_\ell \leq \epsilon_\ell$ *with probability exceeding* $1 - e^{-t}$.

Remark that when $\mathbf{\Omega}$ is in general position ℓ cannot be greater than d and thus p cannot be greater than $2d$ [18]. For this reason, if we want to have large values for p we should allow linear dependencies between the rows of $\mathbf{\Omega}$. In this case the cosparsity of the signal can be greater than d. This explains why linear dependencies are a favorable thing in analysis dictionaries [25]. In Section 7 we shall see that also empirically we get a better recovery when $\mathbf{\Omega}$ contains linear dependencies.

4. Near Optimal Projection

As we will see hereafter, in the proposed algorithms we will face the following problem: Given a general vector $\mathbf{z} \in \mathbb{R}^d$, we would like to find an ℓ-cosparse vector that is closest to it in the ℓ_2-norm sense. In other words, we would like to project the vector to the closest ℓ-cosparse subspace. Given the cosupport Λ of this space the solution is simply $\mathbf{Q}_\Lambda \mathbf{z}$. Thus, the problem of finding the closest ℓ-cosparse vector turns to be the problem of finding the cosupport of the closest ℓ-cosparse subspace. We denote the procedure of finding this cosupport by

$$\mathcal{S}^*_\ell(\mathbf{z}) = \underset{\Lambda \in \mathcal{L}_\ell}{\operatorname{argmin}} \|\mathbf{z} - \mathbf{Q}_\Lambda \mathbf{z}\|_2^2 . \tag{22}$$

In the representation domain in the synthesis case, the support of the closest k-sparse subspace is found simply by hard thresholding, i.e., taking the support of the k-largest elements. However, in the analysis case calculating (22) is NP-complete with no efficient method for doing it for a general $\mathbf{\Omega}$ [33]. Thus an approximation procedure $\hat{\mathcal{S}}_\ell$ is needed. For this purpose we introduce the definition of a near-optimal projection [27].

Definition 4.1. *A procedure $\hat{\mathcal{S}}_\ell$ implies a near-optimal projection $\mathbf{Q}_{\hat{\mathcal{S}}_\ell(\cdot)}$ with a constant C_ℓ if for any $\mathbf{z} \in \mathbb{R}^d$*

$$\left\|\mathbf{z} - \mathbf{Q}_{\hat{\mathcal{S}}_\ell(\mathbf{z})}\mathbf{z}\right\|_2^2 \leq C_\ell \left\|\mathbf{z} - \mathbf{Q}_{\mathcal{S}^*_\ell(\mathbf{z})}\mathbf{z}\right\|_2^2 . \tag{23}$$

A clear implication of this definition is that if $\hat{\mathcal{S}}_\ell$ implies a near-optimal projection with a constant C_ℓ then for any vector $\mathbf{z} \in \mathbb{R}^d$ and an ℓ-cosparse vector $\mathbf{v} \in \mathbb{R}^d$

$$\left\|\mathbf{z} - \mathbf{Q}_{\hat{\mathcal{S}}_\ell(\mathbf{z})}\mathbf{z}\right\|_2^2 \leq C_\ell \|\mathbf{z} - \mathbf{v}\|_2^2 . \tag{24}$$

Similarly to the $\mathbf{\Omega}$-RIP, the above discussion can be directed also for finding the closest vector with corank r defining $\mathcal{S}^{\text{corank}*}_r$ and near optimal projection for this case in a very similar way to (22) and Definition 4.1 respectively.

Having a near-optimal cosupport selection scheme for a general operator is still an open problem and we leave it for a future work. It is possible that this is also NP-complete. We start by describing a simple thresholding rule that can be used with any operator. Even though it does not have any known (near) optimality guarantee besides the case of unitary operators, the numerical section will show it performs well in practice. Then we present two tractable algorithms for finding the optimal cosupport for two non-trivial analysis operators, the one dimensional finite difference operator $\mathbf{\Omega}_{\text{1D-DIF}}$ [34] and the fused Lasso operator $\mathbf{\Omega}_{\text{FUS}}$ [35].

Later in the paper, we propose theoretical guarantees for algorithms that use operators that has an optimal or a near-optimal cosupport selection scheme. We leave the theoretical study of the thresholding technique for a future work but demonstrate its performance empirically in Section 7 where this rule is used showing that also when near-optimality is not at hand reconstruction is feasible.

4.1. Cosupport Selection by Thresholding

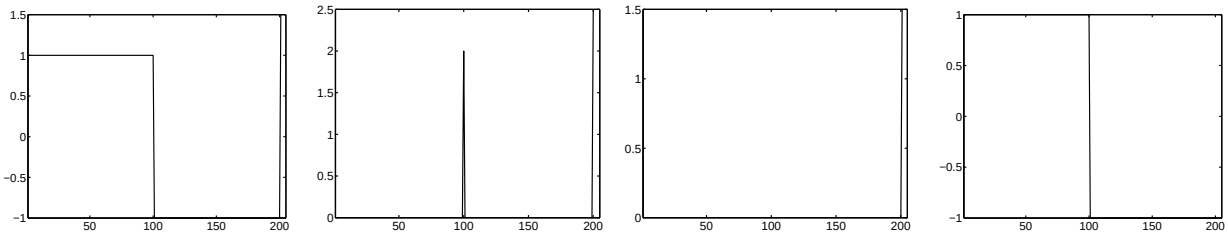
One intuitive option for cosupport selection is the simple thresholding

$$\hat{\mathcal{S}}_\ell(\mathbf{z}) = \text{cosupp}(\mathbf{\Omega}\mathbf{z}, \ell), \tag{25}$$

which selects as a cosupport the indices of the ℓ-smallest elements after applying $\mathbf{\Omega}$ on $\mathbf{z}$. As mentioned above, this selection method is optimal for unitary analysis operators where it coincides with the hard thresholding used in synthesis. However, in the general case this selection method is not guaranteed to give the optimal cosupport. Its near optimality constant C_ℓ is not close to one and is equal to the fraction of the largest and smallest eigenvalues (which are not zero) of the submatrices composed of ℓ rows from $\mathbf{\Omega}$ [27].

One example for an operator for which the thresholding is sub-optimal is the 1D-finite difference operator $\mathbf{\Omega}_{\text{1D-DIF}}$. This operator is defined as:

$$\mathbf{\Omega}_{\text{1D-DIF}} = \begin{pmatrix} -1 & 1 & \cdots & & \\ \vdots & -1 & 1 & & \\ & & & \ddots & \\ & & & -1 & 1 \end{pmatrix} \tag{26}$$



(a) The signal $\mathbf{z}$. (b) $\mathbf{\Omega}_{\text{1D-DIF}}\mathbf{z}$. (c) Projection using thresholding cosupport selection. The projection ℓ_2-norm error is $\sqrt{200}$. (d) Optimal projection. The projection ℓ_2-norm error is 2.5.

Figure 1: Comparison between projection using thresholding cosupport selection and optimal cosupport selection. As it can be seen the thresholding projection error is much larger than the optimal projection error by a factor much larger than 1

In this case, given a signal $\mathbf{z}$, applying $\mathbf{\Omega}_{\text{1D-DIF}}$ on it, result with a vector of coefficients that represents the differences in the signal. The thresholding selection method will select the indices of the ℓ smallest elements in $\mathbf{\Omega z}$ as the cosupport $\Lambda_{\mathbf{z}}$. For example, for the signal $\mathbf{z} \in \mathbb{R}^{201}$ in Fig 1(a) that contains 100 times one, 100 times minus one and 1.5 as the last element, the thresholding will select the cosupport to be the first 199 coefficients in $\mathbf{\Omega}_{\text{1D-DIF}}\mathbf{z}$ that appears in Fig 1(b) and thus the projected vector will be the one in Fig 1(c). Its error in the ℓ_2-norm sense is $\sqrt{200}$. However, selecting the cosupport to be the first 99 elements and last 100 elements result with the projected vector in Fig. 1(d), which has a smaller projection error (2.5). Thus, it is clear that the thresholding is sub-optimal for $\mathbf{\Omega}_{\text{1D-DIF}}$. In a similar way it is also sub-optimal for the 2D-finite difference operator $\mathbf{\Omega}_{\text{2D-DIF}}$ that returns the vertical and horizontal differences of a two dimensional signal. Though not optimal, the use of thresholding with this operator is illustrated in Section 7 demonstrating that also when a good projection is not at hand, good reconstruction is still possible.

4.2. Optimal Analysis Projection Operators

As mentioned above, in general it would appear that determining the optimal projection is computationally difficult with the only general solution being to fully enumerate the projections onto all possible cosupports. Here we highlight two cases where it is relatively easy (polynomial complexity) to calculate the optimal cosparse projection.

4.2.1. Case 1: 1D finite difference

For the 1D finite difference operator the analysis operator is not redundant ($p = d - 1$) but neither is it invertible. As we have seen, a simple thresholding does not provide us with the optimal cosparse projection. Thus, in order to determine the best ℓ-cosparse approximation for a given vector $\mathbf{z}$ we take another route and note that we are looking for the closest (in the ℓ_2-norm sense to $\mathbf{z}$) piecewise constant vector with $p - \ell$ change-points. This problem has been solved previously in the signal processing literature using dynamic programming (DP), see for example: [34]. Thus for this operator it is possible to calculate the best cosparse representation in $O(d^2)$ operations. The existence of a DP solution follows from the ordered localized nature of the finite difference operator. To the best of our knowledge, there is no known extension to 2D finite difference.

4.2.2. Case 2: Fused Lasso Operator

A redundant operator related to the 1D finite difference operator is the so-called fused Lasso operator, usually used with the analysis ℓ_1-minimization [35]. This usually takes the form:

$$\mathbf{\Omega}_{\text{FUS}} = \begin{pmatrix} \mathbf{\Omega}_{\text{1D-DIF}} \\ \epsilon\mathbf{I} \end{pmatrix}. \tag{27}$$

Like $\mathbf{\Omega}_{\text{1D-DIF}}$ this operator works locally and therefore we can expect to derive a DP solution to the approximation problem. This is presented below.

Remark 4.2. *Note that in terms of the cosparsity model the ϵ parameter plays no role. This is in contrast to the traditional convex optimization solutions where the value of ϵ is pivotal [22]. It is possible to mimic the ϵ dependence within the cosparsity framework by considering a generalized fused Lasso operator of the form:*

$$\mathbf{\Omega}_{\epsilon FUS} = \begin{pmatrix} \mathbf{\Omega}_{1D\text{-}DIF} \\ \mathbf{\Omega}_{1D\text{-}DIF} \\ \vdots \\ \mathbf{\Omega}_{1D\text{-}DIF} \\ \mathbf{I} \end{pmatrix}. \tag{28}$$

where the number of repetitions of the $\mathbf{\Omega}_{1D\text{-}DIF}$ operator (and possibly the $\mathbf{I}$ operator) can be selected to mimic a weight on the number of nonzero coefficients of each type. For simplicity we only consider the case indicated by (27)

4.2.3. A recursive solution to the optimal projector for $\mathbf{\Omega}_{FUS}$

Rather than working directly with the operator $\mathbf{\Omega}_{\text{FUS}}$ we make use of the following observation. An ℓ-cosparse vector $\mathbf{v}$ (or k-sparse vector) for $\mathbf{\Omega}_{\text{FUS}}$ is a piecewise constant vector with k_1 change points and k_2 non-zero entries such that $k_1 + k_2 = k = p - \ell$, where $p = 2d - 1$. To understand better the relation between k_1 and k_2, notice that $k_1 = 0$ implies equality of all entries, so $k_2 = 0$ or d, hence $\ell = p$ or $d - 1$. Conversely, considering $d \leq \ell < p$ or $0 \leq \ell < d - 1$ implies $k_1 \neq 0$. It also implies that there is at least one nonzero value, hence $k_2 \neq 0$.

Thus, an ℓ-cosparse vector $\mathbf{v}$ for $\mathbf{\Omega}_{\text{FUS}}$ can be parameterized in terms of a set of change points, $\{n_i\}_{i=0:k_1+1}$, and a set of constants, $\{\mu_i\}_{i=1:k_1+1}$, such that:

$$\mathbf{v}_j = \mu_i, n_{i-1} < j \leq n_i \tag{29}$$

with the convention that $n_0 = 0$ and $n_{k_1+1} = d$, unless stated otherwise. We will also make use of the indicator vector, $\mathbf{s}$, defined as:

$$\mathbf{s}_i = \begin{cases} 0 & \text{if } \mu_i = 0, \\ 1 & \text{otherwise} \end{cases} \text{ for } 1 \leq i \leq k_1 + 1. \tag{30}$$

Using this alternative parametrization we can write the minimum distance between a vector $\mathbf{z}$ and the set of k-sparse fused Lasso coefficients as:

$$\begin{aligned} F_k(\mathbf{z}) = \min_{1 \leq k_1 \leq k} \min_{\substack{\{n_i\}_{i=1:k_1} \\ \{\mu_i\}_{i=1:k_1+1} \\ n_{k_1} < d}} & \sum_{i=1}^{k_1+1} \sum_{j=n_{i-1}+1}^{n_i} (\mathbf{z}_j - \mu_i)^2, \\ \text{subject to } & \sum_{i=1}^{k_1+1} \mathbf{s}_i (n_i - n_{i-1}) = k - k_1 \end{aligned} \tag{31}$$

Although this looks a formidable optimization task we now show that it can be computed recursively through a standard DP strategy, modifying the arguments in [34].

Let us define the optimal cost, $I_k(L, \omega, k_1)$, for the vector $[\mathbf{z}_1, \ldots, \mathbf{z}_L]^T$ with k_1 change points and $\mathbf{s}_{k_1+1} = \omega$, as:

$$\begin{aligned} I_k(L, \omega, k_1) = & \min_{\substack{\{n_i\}_{i=1:k_1} \\ \{\mathbf{s}_i\}_{i=1:k_1+1} \\ n_{k_1} < L,\ n_{k_1+1} = L \\ \mathbf{s}_{k_1+1} = \omega}} \sum_{i=1}^{k_1+1} \sum_{j=n_{i-1}+1}^{n_i} (\mathbf{z}_j - \mu_i)^2, \\ & \text{subject to } \sum_{i=1}^{k_1+1} \mathbf{s}_i (n_i - n_{i-1}) = k - k_1 \\ & \text{and } \mu_i = \frac{\mathbf{s}_i}{n_i - n_{i-1}} \sum_{l=n_{i-1}+1}^{n_i} \mathbf{z}_l \end{aligned} \tag{32}$$

where we have set μ_i to the optimal sample means. Notice that calculating $I_k(L, \omega, k_1)$ is easy for $k_1 \leq k \leq 1$. Thus, we calculate it recursively considering two separate scenarios:

Case 1: $\omega = 0$ where the last block of coefficients are zero. This gives:

$$\begin{aligned} I_k(L,0,k_1) = \min_{n_{k_1}<L} & \left(\sum_{j=n_{k_1}+1}^{L} (\mathbf{z}_j)^2 + \min_{\substack{\{n_i\}_{i=1:k_1-1} \\ \{\mathbf{s}_i\}_{i=1:k_1-1} \\ n_{k_1-1}<n_{k_1} \\ \mathbf{s}_{k_1}=1}} \sum_{i=1}^{k_1} \sum_{j=n_{i-1}+1}^{n_i} (\mathbf{z}_j - \mu_i)^2 \right), \\ & \text{subject to } \sum_{i=1}^{k_1} s_i(n_i - n_{i-1}) = (k-1) - (k_1 - 1) \\ & \text{and } \mu_i = \frac{\mathbf{s}_i}{n_i - n_{i-1}} \sum_{l=n_{i-1}+1}^{n_i} \mathbf{z}_l, \end{aligned} \tag{33}$$

(noting that if $\mathbf{s}_{k_1+1} = 0$ then $\mathbf{s}_{k_1} = 1$ since otherwise n_{k_1} would not have been a change point). This simplifies to the recursive formula:

$$I_k(L,0,k_1) = \min_{n_{k_1}<L} \left(\sum_{j=n_{k_1}+1}^{L} (\mathbf{z}_j)^2 + I_{k-1}(n_{k_1}, 1, k_1 - 1) \right) \tag{34}$$

Case 2: $\omega = 1$ when the final block of coefficients are non-zero we have:

$$\begin{aligned} I_k(L,1,k_1) = \min_{\substack{n_{k_1}<L \\ n_{k_1+1}=L \\ \mathbf{s}_{k_1}}} & \left(\sum_{j=n_{k_1}+1}^{L} (\mathbf{z}_j - \mu_{k_1+1})^2 + \min_{\substack{\{n_i\}_{i=1:k_1-1} \\ \{\mathbf{s}_i\}_{i=1:k_1-1} \\ n_{k_1-1}<n_{k_1}}} \sum_{i=1}^{k_1} \sum_{j=n_{i-1}+1}^{n_i} (\mathbf{z}_j - \mu_i)^2 \right), \\ & \text{subject to } \sum_{i=1}^{k_1} \mathbf{s}_i(n_i - n_{i-1}) = (k - L + n_{k_1} - 1) - (k_1 - 1) \\ & \text{and } \mu_i = \frac{\mathbf{s}_i}{n_i - n_{i-1}} \sum_{l=n_{i-1}+1}^{n_i} \mathbf{z}_l. \end{aligned} \tag{35}$$

This simplifies to the recursive relationship:

$$\begin{aligned} I_k(L,1,k_1) = \min_{\substack{n_{k_1}<L \\ \mathbf{s}_{k_1}}} & \left(\sum_{j=n_{k_1}+1}^{L} (\mathbf{z}_j - \mu_{k_1+1})^2 + I_{k-L+n_{k_1}-1}(n_{k_1}, \mathbf{s}_{k_1}, k_1 - 1) \right) \\ & \text{subject to } \mu_{k_1+1} = \sum_{l=n_{k_1}+1}^{L} \mathbf{z}_l/(L - n_{k_1}) \end{aligned} \tag{36}$$

Equations (34) and (36) are sufficient to enable the calculation of the optimal projection in polynomial time,starting with $k_1 \leq k \leq 1$ and recursively evaluating the costs for $k \geq k_1 \geq 1$. Finally, we have $F_k(\mathbf{z}) = \min_{k_1 \leq k, \omega \in \{0,1\}} I_k(d, \omega, k_1)$. The implementation details are left as an exercise for the reader.

5. New Analysis algorithms

5.1. Quick Review of the Greedy-Like Methods

Before we turn to present the analysis versions of the greedy-like techniques we recall their synthesis versions. These use a prior knowledge about the cardinality k and actually aim at approximating a variant of (3)

$$\underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{M}\mathbf{D}\boldsymbol{\alpha}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{\alpha}\|_0 \leq k. \tag{37}$$

For simplicity we shall present the greedy-like pursuits for the case $\mathbf{D} = \mathbf{I}$. In the general case $\mathbf{M}$ should be replaced with $\mathbf{MD}$, $\mathbf{x}$ with $\boldsymbol{\alpha}$ and the reconstruction result should be $\hat{\mathbf{x}} = \mathbf{D}\hat{\boldsymbol{\alpha}}$. In addition, in the algorithms' description we do not specify the stopping criterion. Any standard stopping criterion, like residual's size or relative iteration change, can be used. More details can be found in [11, 12].

IHT and HTP: IHT [13] and HTP [14] are presented in Algorithm 1. Each IHT iteration is composed of two basic steps. The first is a gradient step, with a step size μ_t, in the direction of minimizing $\|\mathbf{y} - \mathbf{Mx}\|_2^2$. The step size can be either constant in all iterations ($\mu^t = \mu$) or changing [36]. The result vector $\mathbf{x}_g$ is not guaranteed to be sparse and thus the second step of IHT projects $\mathbf{x}_g$ to the closest k-sparse subspace by keeping its largest k elements. The HTP takes a different strategy in the projection step. Instead of using a simple projection to the closest k-sparse subspace, HTP selects the vector in this subspace that minimizes $\|\mathbf{y} - \mathbf{Mx}\|_2^2$ [14, 37].

Algorithm 1 Iterative hard thresholding (IHT) and hard thresholding pursuit (HTP)

Require: $k, \mathbf{M}, \mathbf{y}$ where $\mathbf{y} = \mathbf{Mx} + \mathbf{e}$, k is the cardinality of $\mathbf{x}$ and $\mathbf{e}$ is an additive noise.
Ensure: $\hat{\mathbf{x}}_{\text{IHT}}$ or $\hat{\mathbf{x}}_{\text{HTP}}$: k-sparse approximation of $\mathbf{x}$.
  Initialize representation $\hat{\mathbf{x}}^0 = \mathbf{0}$ and set $t = 0$.
  while halting criterion is not satisfied **do**
    $t = t + 1$.
    Perform a gradient step: $\mathbf{x}_g = \hat{\mathbf{x}}^{t-1} + \mu^t \mathbf{M}^*(\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1})$
    Find a new support: $T^t = \text{supp}(\mathbf{x}_g, k)$
    Calculate a new representation: $\hat{\mathbf{x}}^t_{\text{IHT}} = (\mathbf{x}_g)_{T^t}$ for IHT, and $\hat{\mathbf{x}}^t_{\text{HTP}} = \mathbf{M}^\dagger_{T^t}\mathbf{y}$ for HTP.
  end while
  Form the final solution $\hat{\mathbf{x}}_{\text{IHT}} = \hat{\mathbf{x}}^t_{\text{IHT}}$ for IHT and $\hat{\mathbf{x}}_{\text{HTP}} = \hat{\mathbf{x}}^t_{\text{HTP}}$ for HTP.

CoSaMP and SP: CoSaMP [11] and SP [12] are presented in Algorithm 2. The difference between these two techniques is similar to the difference between IHT and HTP. Unlike IHT and HTP, the estimate for the support of $\mathbf{x}$ in each CoSaMP and SP iteration is computed by observing the residual $\mathbf{y}^t_{\text{resid}} = \mathbf{y} - \mathbf{Mx}^t$. In each iteration, CoSaMP and SP extract new support indices from the residual by taking the indices of the largest elements in $\mathbf{M}^*\mathbf{y}^t_{\text{resid}}$. They add the new indices to the estimated support set from the previous iteration creating a new estimated support $\tilde{T}^t$ with cardinality larger than k. Having the updated support, in a similar way to the projection in HTP, an objective aware projection is performed resulting with an estimate $\mathbf{w}$ for $\mathbf{x}$ that is supported on $\tilde{T}^t$. Since we know that $\mathbf{x}$ is k-sparse we want to project $\mathbf{w}$ to a k-sparse subspace. CoSaMP does it by simple hard thresholding like in IHT. SP does it by an objective aware projection similar to HTP.

5.2. Analysis greedy-like methods

Given the synthesis greedy-like pursuits, we would like to define their analysis counterparts. For this task we need to 'translate' each synthesis operation into an analysis one. This gives us a general recipe for converting algorithms between the two schemes. The parallel lines between the schemes are presented in Table 1. Those become more intuitive and clear when we keep in mind that while the synthesis approach focuses on the non-zeros, the analysis concentrates on the zeros.

For clarity we dwell a bit more on the equivalences. For the cosupport selection, as mentioned in Section 4, computing the optimal cosupport is a combinatorial problem and thus the approximation $\hat{\mathcal{S}}_\ell$ is used. Having a selected cosupport Λ, the projection to its corresponding cosparse subspace becomes trivial, given by $\mathbf{Q}_\Lambda$.

Given two vectors $\mathbf{v}_1 \in \mathcal{A}_{\ell_1}$ and $\mathbf{v}_2 \in \mathcal{A}_{\ell_2}$ such that $\Lambda_1 = \text{cosupp}(\boldsymbol{\Omega}\mathbf{v}_1)$ and $\Lambda_2 = \text{cosupp}(\boldsymbol{\Omega}\mathbf{v}_2)$, we know that $|\Lambda_1| \geq \ell_1$ and $|\Lambda_2| \geq \ell_2$. Denoting $T_1 = \text{supp}(\boldsymbol{\Omega}\mathbf{v}_1)$ and $T_2 = \text{supp}(\boldsymbol{\Omega}\mathbf{v}_2)$ it is clear that $\text{supp}(\boldsymbol{\Omega}(\mathbf{v}_1 + \mathbf{v}_1)) \subseteq T_1 \cup T_2$. Noticing that $\text{supp}(\cdot) = \text{cosupp}(\cdot)^C$ it is clear that $|T_1| \leq p - \ell_1$, $|T_2| \leq p - \ell_2$ and $\text{cosupp}(\boldsymbol{\Omega}(\mathbf{v}_1 + \mathbf{v}_2)) \supseteq (T_1 \cup T_2)^C = T_1^C \cap T_2^C = \Lambda_1 \cap \Lambda_2$. From the last equality we can also deduce that $|\Lambda_1 \cap \Lambda_2| = p - |T_1 \cup T_2| \geq p - (p - \ell_1) - (p - \ell_2) = \ell_1 + \ell_2 - p$.

With the above observations we can develop the analysis versions of the greedy-like algorithms. As in the synthesis case, we do not specify a stopping criterion. Any stopping criterion used for the synthesis versions can be used also for the analysis ones.

Algorithm 2 Subspace Pursuit (SP) and CoSaMP

Require: $k, \mathbf{M}, \mathbf{y}$ where $\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{e}$, k is the cardinality of $\mathbf{x}$ and $\mathbf{e}$ is an additive noise. $a = 1$ (SP), $a = 2$ (CoSaMP).
Ensure: $\hat{\mathbf{x}}_{\text{CoSaMP}}$ or $\hat{\mathbf{x}}_{\text{SP}}$: k-sparse approximation of $\mathbf{x}$.

Initialize the support $T^0 = \emptyset$, the residual $\mathbf{y}^0_{\text{resid}} = \mathbf{y}$ and set $t = 0$.
while halting criterion is not satisfied **do**
  $t = t + 1$.
  Find new support elements: $T_\Delta = \text{supp}(\mathbf{M}^* \mathbf{y}^{t-1}_{\text{resid}}, ak)$.
  Update the support: $\tilde{T}^t = T^{t-1} \cup T_\Delta$.
  Compute a temporary representation: $\mathbf{w} = \mathbf{M}^\dagger_{\tilde{T}^t} \mathbf{y}$.
  Prune small entries: $T^t = \text{supp}(\mathbf{w}, k)$.
  Calculate a new representation: $\hat{\mathbf{x}}^t_{\text{CoSaMP}} = \mathbf{w}_{T^t}$ for CoSaMP, and $\hat{\mathbf{x}}^t_{\text{SP}} = \mathbf{M}^\dagger_{T^t} \mathbf{y}$ for SP.
  Update the residual: $\mathbf{y}^t_{\text{resid}} = \mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t_{\text{CoSaMP}}$ for CoSaMP, and $\mathbf{y}^t_{\text{resid}} = \mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t_{\text{SP}}$ for SP.
end while
Form the final solution $\hat{\mathbf{x}}_{\text{CoSaMP}} = \hat{\mathbf{x}}^t_{\text{CoSaMP}}$ for CoSaMP and $\hat{\mathbf{x}}_{\text{SP}} = \hat{\mathbf{x}}^t_{\text{SP}}$ for SP.

Synthesis operation name	**Synthesis operation**	**Analysis operation name**	**Analysis operation**
Support selection	Largest k elements: $T = \text{supp}(\cdot, k)$	Cosupport selection	Using a near optimal projection: $\Lambda = \hat{\mathcal{S}}_\ell(\cdot)$
Orthogonal Projection of $\mathbf{z}$ to a k-sparse subspace with support T	$\mathbf{z}_T$	Orthogonal projection of $\mathbf{z}$ to an ℓ-cosparse subspace with cosupport Λ	$\mathbf{Q}_\Lambda \mathbf{z}$
Objective aware projection to a k-sparse subspace with support T	$\mathbf{M}^\dagger_T \mathbf{y} =$ $\text{argmin}_{\mathbf{v}} \|\mathbf{y} - \mathbf{M}\mathbf{v}\|_2^2$ s.t. $\mathbf{v}_{T^C} = 0$	Objective aware projection to an ℓ-cosparse subspace with cosupport Λ	$\text{argmin}_{\mathbf{v}} \|\mathbf{y} - \mathbf{M}\mathbf{v}\|_2^2$ s.t. $\mathbf{\Omega}_\Lambda \mathbf{v} = 0$
Support of $\mathbf{v}_1 + \mathbf{v}_2$ where $\text{supp}(\mathbf{v}_1) = T_1$ and $\text{supp}(\mathbf{v}_2) = T_2$	$\text{supp}(\mathbf{v}_1 + \mathbf{v}_2) \subseteq T_1 \cup T_2$	Cosupport of $\mathbf{v}_1 + \mathbf{v}_2$ where $\text{cosupp}(\mathbf{v}_1) = \Lambda_1$ and $\text{cosupp}(\mathbf{v}_2) = \Lambda_2$	$\text{cosupp}(\mathbf{v}_1 + \mathbf{v}_2) \supseteq \Lambda_1 \cap \Lambda_2$
Maximal size of $T_1 \cup T_2$ where $\|T_1\| \le k_1$ and $\|T_2\| \le k_2$	$\|T_1 \cup T_2\| \le k_1 + k_2$	Minimal size of $\Lambda_1 \cap \Lambda_2$ where $\|\Lambda_1\| \ge \ell_1$ and $\|\Lambda_2\| \ge \ell_2$	$\|\Lambda_1 \cap \Lambda_2\| \ge \ell_1 + \ell_2 - p$

Table 1: Parallel synthesis and analysis operations

AIHT and AHTP: Analysis IHT (AIHT) and analysis HTP (AHTP) are presented in Algorithm 3. As in the synthesis case, the choice of the gradient stepsize μ^t is crucial: If μ^t's are chosen too small, the algorithm gets stuck at a wrong solution and if too large, the algorithm diverges. We consider two options for μ^t.

In the first we choose $\mu^t = \mu$ for some constant μ for all iterations. A theoretical discussion on how to choose μ properly is given in Section 6.1.

The second option is to select a different μ in each iteration. One way for doing it is to choose an 'optimal' stepsize μ^t by solving the following problem

$$\mu^t := \underset{\mu}{\text{argmin}} \left\| \mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t \right\|_2^2 . \tag{38}$$

Since $\hat{\Lambda}^t = \hat{\mathcal{S}}_\ell(\hat{\mathbf{x}}^{t-1} + \mu^t \mathbf{M}^*(\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}))$ and $\hat{\mathbf{x}}^t = \mathbf{Q}_{\hat{\Lambda}^t}(\mathbf{x}_g)$, the above requires a line search over different values of μ and along the search $\hat{\Lambda}^t$ might change several times. A simpler way is an adaptive step size selection as proposed in [36] for IHT. In a heuristical way we limit the search to the cosupport $\tilde{\Lambda} = \hat{\mathcal{S}}_\ell(\mathbf{M}^*(\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1})) \cap \hat{\Lambda}^{t-1}$. This is the intersection of the cosupport of $\hat{\mathbf{x}}^{t-1}$ with the ℓ-cosparse cosupport of the estimated closest ℓ-cosparse subspace to $\mathbf{M}^*(\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1})$. Since $\hat{\mathbf{x}}^{t-1} = \mathbf{Q}_{\tilde{\Lambda}} \hat{\mathbf{x}}^{t-1}$, finding μ turns to be

$$\mu^t := \underset{\mu}{\text{argmin}} \left\| \mathbf{y} - \mathbf{M}(\hat{\mathbf{x}}^{t-1} + \mu \mathbf{Q}_{\tilde{\Lambda}} \mathbf{M}^*(\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1})) \right\|_2^2 . \tag{39}$$

Algorithm 3 Analysis Iterative hard thresholding (AIHT) and analysis hard thresholding pursuit (AHTP)

Require: $\ell, \mathbf{M}, \mathbf{\Omega}, \mathbf{y}$ where $\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{e}$, ℓ is the cosparsity of $\mathbf{x}$ under $\mathbf{\Omega}$ and $\mathbf{e}$ is the additive noise.
Ensure: $\hat{\mathbf{x}}_{\text{AIHT}}$ or $\hat{\mathbf{x}}_{\text{AHTP}}$: ℓ-cosparse approximation of $\mathbf{x}$.
Initialize estimate $\hat{\mathbf{x}}^0 = \mathbf{0}$ and set $t = 0$.
while halting criterion is not satisfied **do**
$t = t + 1$.
Perform a gradient step: $\mathbf{x}_g = \hat{\mathbf{x}}^{t-1} + \mu^t \mathbf{M}^*(\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1})$
Find a new cosupport: $\hat{\Lambda}^t = \hat{\mathcal{S}}_\ell(\mathbf{x}_g)$
Calculate a new estimate: $\hat{\mathbf{x}}^t_{\text{AIHT}} = \mathbf{Q}_{\hat{\Lambda}^t}\mathbf{x}_g$ for AIHT, and $\hat{\mathbf{x}}^t_{\text{AHTP}} = \text{argmin}_{\tilde{\mathbf{x}}} \|\mathbf{y} - \mathbf{M}\tilde{\mathbf{x}}\|_2^2$ s.t. $\mathbf{\Omega}_{\hat{\Lambda}^t}\tilde{\mathbf{x}} = 0$ for AHTP.
end while
Form the final solution $\hat{\mathbf{x}}_{\text{AIHT}} = \hat{\mathbf{x}}^t_{\text{AIHT}}$ for AIHT and $\hat{\mathbf{x}}_{\text{AHTP}} = \hat{\mathbf{x}}^t_{\text{AHTP}}$ for AHTP.

Algorithm 4 Analysis Subspace Pursuit (ASP) and Analysis CoSaMP (ACoSaMP)

Require: $\ell, \mathbf{M}, \mathbf{\Omega}, \mathbf{y}, a$ where $\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{e}$, ℓ is the cosparsity of $\mathbf{x}$ under $\mathbf{\Omega}$ and $\mathbf{e}$ is the additive noise.
Ensure: $\hat{\mathbf{x}}_{\text{ACoSaMP}}$ or $\hat{\mathbf{x}}_{\text{ASP}}$: ℓ-cosparse approximation of $\mathbf{x}$.
Initialize the cosupport $\Lambda^0 = \{i, 1 \le i \le p\}$, the residual $\mathbf{y}^0_{\text{resid}} = \mathbf{y}$ and set $t = 0$.
while halting criterion is not satisfied **do**
$t = t + 1$.
Find new cosupport elements: $\Lambda_\Delta = \hat{\mathcal{S}}_{a\ell}(\mathbf{M}^*\mathbf{y}^{t-1}_{\text{resid}})$.
Update the cosupport: $\tilde{\Lambda}^t = \hat{\Lambda}^{t-1} \cap \Lambda_\Delta$.
Compute a temporary estimate: $\mathbf{w} = \text{argmin}_{\tilde{\mathbf{x}}} \|\mathbf{y} - \mathbf{M}\tilde{\mathbf{x}}\|_2^2$ s.t. $\mathbf{\Omega}_{\tilde{\Lambda}^t}\tilde{\mathbf{x}} = 0$.
Enlarge the cosupport: $\hat{\Lambda}^t = \hat{\mathcal{S}}_\ell(\mathbf{w})$.
Calculate a new estimate: $\hat{\mathbf{x}}^t_{\text{ACoSaMP}} = \mathbf{Q}_{\hat{\Lambda}^t}\mathbf{w}$ for ACoSaMP, and $\hat{\mathbf{x}}^t_{\text{ASP}} = \text{argmin}_{\tilde{\mathbf{x}}} \|\mathbf{y} - \mathbf{M}\tilde{\mathbf{x}}\|_2^2$ s.t. $\mathbf{\Omega}_{\hat{\Lambda}^t}\tilde{\mathbf{x}} = 0$ for ASP.
Update the residual: $\mathbf{y}^t_{\text{resid}} = \mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t_{\text{ACoSaMP}}$ for ACoSaMP, and $\mathbf{y}^t_{\text{resid}} = \mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t_{\text{ASP}}$ for ASP.
end while
Form the final solution $\hat{\mathbf{x}}_{\text{ACoSaMP}} = \hat{\mathbf{x}}^t_{\text{ACoSaMP}}$ for ACoSaMP and $\hat{\mathbf{x}}_{\text{ASP}} = \hat{\mathbf{x}}^t_{\text{ASP}}$ for ASP.

This procedure of selecting μ^t does not require a line search and it has a simple closed form solution.

To summarize, there are three main options for the step size selection:

- Constant step-size selection – uses a constant step size $\mu^t = \mu$ in all iterations.
- Optimal changing step-size selection – uses different values for μ^t in each iterations by minimizing $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2$.
- Adaptive changing step-size selection – uses (39).

ACoSaMP and ASP: analysis CoSaMP (ACoSaMP) and analysis SP (ASP) are presented in Algorithm 4. The stages are parallel to those of the synthesis CoSaMP and SP. We dwell a bit more on the meaning of the parameter a in the algorithms. This parameter determines the size of the new cosupport Λ_Δ in each iteration. $a = 1$ means that the size is ℓ and according to Table 1 it is equivalent to $a = 1$ in the synthesis as done in SP in which we select new k indices for the support in each iteration. In synthesis CoSaMP we use $a = 2$ and select $2k$ new elements. $2k$ is the maximal support size of two added k-sparse vectors. The corresponding minimal size in the analysis case is $2\ell - p$ according to Table 1. For this setting we need to choose $a = \frac{2\ell - p}{\ell}$.

5.3. The Unitary Case

For $\mathbf{\Omega} = \mathbf{I}$ the synthesis and the analysis greedy-like algorithms become equivalent. This is easy to see since in this case we have $p = d$, $k = d - \ell$, $\Lambda = T^C$, $\mathbf{Q}_\Lambda\mathbf{x} = \mathbf{x}_T$ and $T_1 \cup T_2 = \Lambda_1 \cap \Lambda_2$ for $\Lambda_1 = T_1^C$ and $\Lambda_2 = T_2^C$. In addition, $\hat{\mathcal{S}}_\ell = \mathcal{S}^*_\ell$ finds the closest ℓ-cosparse subspace by simply taking the smallest ℓ elements. Using similar arguments, also in the case where $\mathbf{\Omega}$ is a unitary matrix the analysis methods coincide with the synthesis ones. In order to get exactly the same algorithms $\mathbf{M}$ is replaced with $\mathbf{M}\mathbf{\Omega}^*$ in the synthesis techniques and the output is multiplied by $\mathbf{\Omega}^*$.

Based on this observation, we can deduce that the guarantees of the synthesis greedy-like methods apply also for the analysis ones in a trivial way. Thus, it is tempting to assume that the last should have similar guarantees based on the $\mathbf{\Omega}$-RIP. In the next section we develop such claims.

5.4. Relaxed Versions for High Dimensional Problems

Before moving to the next section we mention a variation of the analysis greedy-like techniques. In AHTP, ACoSaMP and ASP we need to solve the constrained minimization problem $\min_{\tilde{\mathbf{x}}} \|\mathbf{y}-\mathbf{M}\tilde{\mathbf{x}}\|_2^2$ s.t. $\|\mathbf{\Omega}_\Lambda \tilde{\mathbf{x}}\|_2^2 = 0$. For high dimensional signals this problem is hard to solve and we suggest to replace it with minimizing $\|\mathbf{y}-\mathbf{M}\tilde{\mathbf{x}}\|_2^2 + \lambda \|\mathbf{\Omega}_\Lambda \tilde{\mathbf{x}}\|_2^2$, where λ is a relaxation constant. This results in a relaxed version of the algorithms. We refer hereafter to these versions as relaxed AHTP (RAHTP) relaxed ASP (RASP) and relaxed ACoSaMP (RACoSaMP).

6. Algorithms Guarantees

In this section we provide theoretical guarantees for the reconstruction performance of the analysis greedy-like methods. For AIHT and AHTP we study both the constant step-size and the optimal step-size selections. For ACoSaMP and ASP the analysis is made for $a = \frac{2\ell-p}{\ell}$, but we believe that it can be extended also to other values of a, such as $a = 1$. The performance guarantees we provide are summarized in the following two theorems. The first theorem, for AIHT and AHTP, is a simplified version of Theorem 6.5 and the second theorem, for ASP and ACoSaMP, is a combination of Corollaries 6.9 and 6.14, all of which appear hereafter along with their proofs. Before presenting the theorems we recall the problem we aim at solving:

Definition 6.1 (Problem $\mathcal{P}$). *Consider a measurement vector $\mathbf{y} \in \mathbb{R}^m$ such that $\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{e}$ where $\mathbf{x} \in \mathbb{R}^d$ is ℓ-cosparse, $\mathbf{M} \in \mathbb{R}^{m\times d}$ is a degradation operator and $\mathbf{e} \in \mathbb{R}^m$ is a bounded additive noise. The largest singular value of $\mathbf{M}$ is $\sigma_{\mathbf{M}}$ and its $\mathbf{\Omega}$-RIP constant is δ_ℓ. The analysis operator $\mathbf{\Omega} \in \mathbb{R}^{p\times d}$ is given and fixed. A procedure $\hat{\mathcal{S}}_\ell$ for finding a cosupport that implies a near optimal projection with a constant C_ℓ is assumed to be at hand. Our task is to recover $\mathbf{x}$ from $\mathbf{y}$. The recovery result is denoted by $\hat{\mathbf{x}}$.*

Theorem 6.2 (Stable Recovery of AIHT and AHTP). *Consider the problem $\mathcal{P}$ and apply either AIHT or AHTP with a certain constant step-size or an optimal changing step-size, obtaining $\hat{\mathbf{x}}^t$ after t iterations. If*

$$\frac{(C_\ell - 1)\sigma_{\mathbf{M}}^2}{C_\ell} < 1 \tag{40}$$

and

$$\delta_{2\ell-p} < \delta_1(C_\ell, \sigma_{\mathbf{M}}^2),$$

where $\delta_1(C_\ell, \sigma_{\mathbf{M}}^2)$ is a constant guaranteed to be greater than zero whenever (40) *is satisfied and C_ℓ is the near-optimal projection constant for cosparsity ℓ (Definition 4.1), then after a finite number of iterations t^**

$$\left\|\mathbf{x} - \hat{\mathbf{x}}^{t^*}\right\|_2 \le c_1 \|\mathbf{e}\|_2 , \tag{41}$$

implying that these algorithms lead to a stable recovery. The constant c_1 is a function of $\delta_{2\ell-p}$, C_ℓ and $\sigma_{\mathbf{M}}^2$, and the constant step-size used is dependent on $\delta_1(C_\ell, \sigma_{\mathbf{M}}^2)$.

Theorem 6.3 (Stable Recovery of ASP and ACoSaMP). *Consider the problem $\mathcal{P}$ and apply either ACoSaMP or ASP with $a = \frac{2\ell-p}{\ell}$, obtaining $\hat{\mathbf{x}}^t$ after t iterations. If*

$$\frac{(C_{\hat{\mathcal{S}}}^2 - 1)\sigma_{\mathbf{M}}^2}{C_{\hat{\mathcal{S}}}^2} < 1, \tag{42}$$

and

$$\delta_{4\ell-3p} < \delta_2(C_{\hat{\mathcal{S}}}, \sigma_{\mathbf{M}}^2),$$

where $C_{\hat{\mathcal{S}}} = \max(C_\ell, C_{2\ell-p})$ *and* $\delta_2(C_{\hat{\mathcal{S}}}, \sigma_{\mathbf{M}}^2)$ *is a constant guaranteed to be greater than zero whenever* (42) *is satisfied, then after a finite number of iterations* t^*

$$\left\|\mathbf{x} - \hat{\mathbf{x}}^{t^*}\right\|_2 \leq c_2 \|\mathbf{e}\|_2 , \tag{43}$$

implying that these algorithms lead to a stable recovery. The constant c_2 *is a function of* $\delta_{4\ell-3p}$, C_ℓ, $C_{2\ell-p}$ *and* $\sigma_{\mathbf{M}}^2$.

Before we proceed to the proofs, let us comment on the constants in the above theorems. Their values can be calculated using Theorem 6.5, and Corollaries 6.9 and 6.14. In the case where $\mathbf{\Omega}$ is a unitary matrix, (40) and (42) are trivially satisfied since $C_\ell = C_{2\ell-p} = 1$. In this case the $\mathbf{\Omega}$-RIP conditions become $\delta_{2\ell-p} < \delta_1(1, \sigma_{\mathbf{M}}^2) = 1/3$ for AIHT and AHTP, and $\delta_{4\ell-3p} < \delta_2(1, \sigma_{\mathbf{M}}^2) = 0.0156$ for ACoSaMP and ASP. In terms of synthesis RIP for $\mathbf{M\Omega}^*$, the condition $\delta_{2\ell-p} < 1/3$ parallels $\delta_{2k}(\mathbf{M\Omega}^*) < 1/3$ and similarly $\delta_{4\ell-3p} < 0.0156$ parallels $\delta_{4k}(\mathbf{M\Omega}^*) < 0.0156$. Note that the condition we pose for AIHT and AHTP in this case is the same as the one presented for synthesis IHT with a constant step size [16]. Better reference constants were achieved in the synthesis case for all four algorithms and thus we believe that there is still room for improvement of the reference constants in the analysis context.

In the non-unitary case, the value of $\sigma_{\mathbf{M}}$ plays a vital role, though we believe that this is just an artifact of our proof technique. For a random Gaussian matrix whose entries are i.i.d with a zero-mean and a variance $\frac{1}{m}$, $\sigma_{\mathbf{M}}$ behaves like $\frac{d}{m}\left(1 + \sqrt{\frac{d}{m}}\right)$. This is true also for other types of distributions for which the fourth moment is known to be bounded [38]. For example, for $d/m = 1.5$ we have found empirically that $\sigma_{\mathbf{M}}^2 \simeq 5$. In this case we need $C_\ell \leq \frac{5}{4}$ for (40) to hold and $C_{\hat{\mathcal{S}}} \leq 1.118$ for (42) to hold, and both are quite demanding on the quality of the near-optimal projection. For $C_\ell = C_{\hat{\mathcal{S}}} = 1.05$ we have the conditions $\delta_{2\ell-p} \leq 0.289$ for AIHT and AHTP, and $\delta_{4\ell-3p} \leq 0.0049$ for ACoSaMP and ASP; and for $C_\ell = C_{\hat{\mathcal{S}}} = 1.1$ we have $\delta_{2\ell-p} \leq 0.24$ for AIHT and AHTP, and $\delta_{4\ell-3p} \leq 0.00032$ for ACoSaMP and ASP.

As in the synthesis case, the $\mathbf{\Omega}$-RIP requirements for the theoretical bounds of AIHT and AHTP are better than those for ACoSaMP and ASP. In addition, in the migration from the synthesis to the analysis we lost more precision in the bounds for ACoSaMP and ASP than in those of AIHT and AHTP. In particular, even in the case where $\mathbf{\Omega}$ is the identity we do not coincide with any of the synthesis parallel RIP reference constants. We should also remember that the synthesis bound for SP is in terms of δ_{3k} and not δ_{4k} [12]. Thus, we expect that it will be possible to give a condition for ASP in terms of $\delta_{3\ell-2p}$ with better reference constants. However, our main interest in this work is to show the existence of such bounds, and in Section 6.5 we dwell more on their meaning.

We should note that here and elsewhere we can replace the conditions on $\delta_{2\ell-p}$ and $\delta_{4\ell-3p}$ in the theorems to conditions on $\delta_{2r-p}^{\text{corank}}$ and $\delta_{4r-3p}^{\text{corank}}$ and the proofs will be almost the same[2]. In this case we will be analyzing a version of the algorithms which is driven by the corank instead of the cosparsity. This would mean we need the near-optimal projection to be in terms of the corank. In the case where $\mathbf{\Omega}$ is in a general position, there is no difference between the cosparsity ℓ and the corank r. However, when we have linear dependencies in $\mathbf{\Omega}$ the two measures differ and an ℓ-cosparse vector is not necessarily a vector with a corank r.

As we will see hereafter, our recovery conditions require $\delta_{2\ell-p}$ and $\delta_{4\ell-3p}$ to be as small as possible and for this we need $2\ell - p$ and $4\ell - 3p$ to be as large as possible. Thus, we need ℓ to be as close as possible to p and for highly redundant $\mathbf{\Omega}$ this cannot be achieved without having linear dependencies in $\mathbf{\Omega}$. Apart from the theoretical advantage of linear dependencies in $\mathbf{\Omega}$, we also show empirically that an analysis dictionary with linear dependencies has better recovery rate than analysis dictionary in a general position of the same dimension. Thus, we deduce that linear dependencies in $\mathbf{\Omega}$ lead to better bounds and restoration performance.

Though linear dependencies allow ℓ to be larger than d and be in the order of p, the value of the corank is always bounded by d and cannot be expected to be large enough for highly redundant analysis dictionaries. In addition, we will see hereafter that the number of measurements m required by the $\mathbf{\Omega}$-RIP is strongly dependent on ℓ and less effected by the value of r. From the computational point of view we note also that using corank requires its computation in each iteration which increases the overall complexity of the algorithms. Thus, it is more reasonable to have conditions on $\delta_{2\ell-p}$ and $\delta_{4\ell-3p}$ than on $\delta_{2r-p}^{\text{corank}}$ and $\delta_{4r-3p}^{\text{corank}}$, and our study will be focused on the cosparsity based algorithms.

[2] At a first glance one would think that the conditions should be in terms of $\delta_{2r-d}^{\text{corank}}$ and $\delta_{4r-3d}^{\text{corank}}$. However, given two cosparse vectors with coranks r_1 and r_2 the best estimation we can have for the corank of their sum is $r_1 + r_2 - p$.

6.1. AIHT and AHTP Guarantees

A uniform guarantee for AIHT in the case that an optimal projection is given, is presented in [29]. The work in [29] dealt with a general union of subspaces, $\mathcal{A}$, and assumed that $\mathbf{M}$ is bi-Lipschitz on the considered union of subspaces. In our case $\mathcal{A} = \mathcal{A}_\ell$ and the bi-Lipschitz constants of $\mathbf{M}$ are the largest B_L and smallest B_U where $0 < B_L \leq B_U$ such that for all ℓ-cosparse vectors $\mathbf{v}_1, \mathbf{v}_2$:

$$B_L \left\|\mathbf{v}_1 + \mathbf{v}_2\right\|_2^2 \leq \left\|\mathbf{M}(\mathbf{v}_1 + \mathbf{v}_2)\right\|_2^2 \leq B_U \left\|\mathbf{v}_1 + \mathbf{v}_2\right\|_2^2 . \tag{44}$$

Under this assumption, one can apply Theorem 2 from [29] to the idealized AIHT that has access to an optimal projection and uses a constant step size $\mu^t = \mu$. Relying on Table 1 we present this theorem and replace B_L and B_U with $1 - \delta_{2\ell-p}$ and $1 + \delta_{2\ell-p}$ respectively.

Theorem 6.4 (Theorem 2 in [29]). *Consider the problem $\mathcal{P}$ with $C_\ell = 1$ and apply AIHT with a constant step size μ. If $1 + \delta_{2\ell-p} \leq \frac{1}{\mu} < 1.5(1 - \delta_{2\ell-p})$ then after a finite number of iterations t^**

$$\left\|\mathbf{x} - \hat{\mathbf{x}}^{t^*}\right\|_2 \leq c_3 \left\|\mathbf{e}\right\|_2 , \tag{45}$$

implying that AIHT leads to a stable recovery. The constant c_3 is a function of $\delta_{2\ell-p}$ and μ.

In this work we extend the above in several ways: First, we refer to the case where optimal projection is not known, and show that the same flavor guarantees apply for a near-optimal projection[3]. The price we seemingly have to pay is that $\sigma_{\mathbf{M}}$ enters the game. Second, we derive similar results for the AHTP method. Finally, we also consider the optimal step size and show that the same performance guarantees hold true in that case.

Theorem 6.5. *Consider the problem $\mathcal{P}$ and apply either AIHT or AHTP with a constant step size μ or an optimal changing step size. For a positive constant $\eta > 0$, let*

$$b_1 := \frac{\eta}{1+\eta} \quad \text{and} \quad b_2 := \frac{(C_\ell - 1)\sigma_{\mathbf{M}}^2 b_1^2}{C_\ell(1 - \delta_{2\ell-p})}.$$

Suppose $\frac{b_2}{b_1^2} = \frac{(C_\ell-1)\sigma_{\mathbf{M}}^2}{C_\ell(1-\delta_{2\ell-p})} < 1$, $1 + \delta_{2\ell-p} \leq \frac{1}{\mu} < \left(1 + \sqrt{1 - \frac{b_2}{b_1^2}}\right) b_1(1 - \delta_{2\ell-p})$ and $\frac{1}{\mu} \leq \sigma_{\mathbf{M}}^2$. Then for

$$t \geq t^* \triangleq \frac{\log\left(\frac{\eta\|\mathbf{e}\|_2^2}{\|\mathbf{y}\|_2^2}\right)}{\log\left((1+\frac{1}{\eta})^2(\frac{1}{\mu(1-\delta_{2\ell-p})} - 1)C_\ell + (C_\ell - 1)(\mu\sigma_{\mathbf{M}}^2 - 1) + \frac{C_\ell}{\eta^2}\right)}, \tag{46}$$

$$\left\|\mathbf{x} - \hat{\mathbf{x}}^t\right\|_2^2 \leq \frac{(1+\eta)^2}{1 - \delta_{2\ell-p}} \left\|\mathbf{e}\right\|_2^2 , \tag{47}$$

implying that AIHT and AHTP lead to a stable recovery. Note that for an optimal changing step-size we set $\mu = \frac{1}{1+\delta_{2\ell-p}}$ in t^* and the theorem conditions turn to be $\frac{b_2}{b_1^2} < 1$ and $1 + \delta_{2\ell-p} < (1 + \sqrt{1 - \frac{b_2}{b_1^2}}) b_1(1 - \delta_{2\ell-p})$.

This theorem is the parallel to Theorems 2.1 in [16] for IHT. A few remarks are in order for the nature of the theorem, especially in regards to the constant η. One can view that η gives a trade-off between satisfying the theorem conditions and the amplification of the noise. In particular, one may consider that the above theorem proves the convergence result for the noiseless case by taking η to infinity; one can imagine solving the problem $\mathcal{P}$ where $\mathbf{e} \to 0$, and applying the theorem with appropriately chosen η which approaches infinity. It is indeed possible to show that the iterate solutions of AIHT and AHTP converges to $\mathbf{x}$ when there is no noise. However, we will not give a separate proof since the basic idea of the arguments is the same for both cases.

[3] Remark that we even improve the condition of the idealized case in [29] to be $\delta_{2\ell-p} \leq \frac{1}{3}$ instead of $\delta_{2\ell-p} \leq \frac{1}{5}$.

As to the minimal number of iterations t^* given in (46), one may ask whether it can be negative. In order to answer this question it should be noted that according to the conditions of the Theorem the term inside the log in the denominator (46) is always greater than zero. Thus, t^* will be negative only if $\|\mathbf{y}\|_2^2 < \eta \|\mathbf{e}\|_2^2$. Indeed, in this case 0 iterations suffice for having the bound in (47).

The last remark is on the step-size selection. The advantage of the optimal changing step-size over the constant step-size is that we get the guarantee of the optimal constant step-size $\mu = \frac{1}{1+\delta_{2\ell-p}}$ without computing it. This is important since in practice we cannot evaluate the value of $\delta_{2\ell-p}$. However, the disadvantage of using the optimal changing step-size is its additional complexity for the algorithm. Thus, one option is to approximate the optimal selection rule by replacing it with an adaptive one, for which we do not have a theoretical guarantee. Another option is to set $\mu = 6/5$ which meets the theorem conditions for small enough $\delta_{2\ell-p}$, in the case where an optimal projection is at hand.

We will prove the theorem by proving two key lemmas first. The proof technique is based on ideas from [16] and [29]. Recall that the two iterative algorithms try to reduce the objective $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2$ over iterations t. Thus, the progress of the algorithms can be indirectly measured by how much the objective $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2$ is reduced at each iteration t. The two lemmas that we present capture this idea. The first lemma is similar to Lemma 3 in [29] and relates $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2$ to $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2$ and similar quantities at iteration $t-1$. We remark that the constraint $\frac{1}{\mu} \leq \sigma_{\mathbf{M}}^2$ in Theorem 6.5 may not be necessary and is added only for having a simpler derivation of the results in this theorem. Furthermore, this is a very mild condition compared to $\frac{1}{\mu} < \left(1 + \sqrt{1 - \frac{b_2}{b_1^2}}\right) b_1(1 - \delta_{2\ell-p})$ and can only limit the range of values that can be used with the constant step size versions of the algorithms.

Lemma 6.6. *Consider the problem $\mathcal{P}$ and apply either AIHT or AHTP with a constant step size μ satisfying $\frac{1}{\mu} \geq 1 + \delta_{2\ell-p}$ or an optimal step size. Then, at the t-th iteration, the following holds:*

$$\begin{aligned}\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2 - \left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2 \leq C_\ell \left(\|\mathbf{y} - \mathbf{M}\mathbf{x}\|_2^2 - \left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2\right) \\ + C_\ell \left(\frac{1}{\mu(1-\delta_{2\ell-p})} - 1\right) \left\|\mathbf{M}(\mathbf{x} - \hat{\mathbf{x}}^{t-1})\right\|_2^2 + (C_\ell - 1)\mu\sigma_{\mathbf{M}}^2 \left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2.\end{aligned} \tag{48}$$

For the optimal step size the bound is achieved with the value $\mu = \frac{1}{1+\delta_{2\ell-p}}$.

The proof of the above lemma appears in Appendix B. The second lemma is built on the result of Lemma 6.6. It shows that once the objective $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2$ at iteration $t-1$ is small enough, then we are guaranteed to have small $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2$ as well. Given the presence of noise, this is quite natural; one cannot expect it to approach 0 but may expect it not to become worse. Moreover, the lemma also shows that if $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2$ is not small, then the objective in iteration t is necessarily reduced by a constant factor.

Lemma 6.7. *Suppose that the same conditions of Theorem 6.5 hold true. If $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2 \leq \eta^2 \|\mathbf{e}\|_2^2$, then $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2 \leq \eta^2 \|\mathbf{e}\|_2^2$. Furthermore, if $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2 > \eta^2 \|\mathbf{e}\|_2^2$, then*

$$\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2 \leq c_4 \left\|\mathbf{y} - \mathbf{M}\mathbf{x}^{t-1}\right\|_2^2 \tag{49}$$

where

$$c_4 := \left(1 + \frac{1}{\eta}\right)^2 \left(\frac{1}{\mu(1-\delta_{2\ell-p})} - 1\right) C_\ell + (C_\ell - 1)(\mu\sigma_{\mathbf{M}}^2 - 1) + \frac{C_\ell}{\eta^2} < 1.$$

Having the two lemmas above, the proof of the theorem is straightforward.

Proof:[Proof of Theorem 6.5] When we initialize $\hat{\mathbf{x}}^0 = \mathbf{0}$, we have $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^0\right\|_2^2 = \|\mathbf{y}\|_2^2$. Assuming that $\|\mathbf{y}\|_2 > \eta \|\mathbf{e}\|_2$ and applying Lemma 6.7 repeatedly, we obtain

$$\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2 \leq \max(c_4^t \|\mathbf{y}\|_2^2, \eta^2 \|\mathbf{e}\|_2^2).$$

Since $c_4^t \|\mathbf{y}\|_2^2 \leq \eta^2 \|\mathbf{e}\|_2^2$ for $t \geq t^*$, we have simply

$$\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2 \leq \eta^2 \|\mathbf{e}\|_2^2 \tag{50}$$

for $t \geq t^*$. If $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^0\right\|_2 = \|\mathbf{y}\|_2 \leq \eta \|\mathbf{e}\|_2$ then according to Lemma 6.7, (50) holds for every $t > 0$. Finally, we observe

$$\left\|\mathbf{x} - \hat{\mathbf{x}}^t\right\|_2^2 \leq \frac{1}{1 - \delta_{2\ell-p}} \left\|\mathbf{M}(\mathbf{x} - \hat{\mathbf{x}}^t)\right\|_2^2 \tag{51}$$

and, by the triangle inequality,

$$\left\|\mathbf{M}(\mathbf{x} - \hat{\mathbf{x}}^t)\right\|_2 \leq \left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2 + \|\mathbf{e}\|_2 . \tag{52}$$

By plugging (50) into (52) and then the resulted inequality into (51), the result of the Theorem follows. □

As we have seen, the above AIHT and AHTP results hold for the cases of using a constant or an optimal changing step size. The advantage of using an optimal one is that we do not need to find μ that satisfies the conditions of the theorem – the knowledge that such a μ exists is enough. However, its disadvantage is the additional computational complexity it introduces. In Section 5 we have introduced a third option of using an approximated adaptive step size. In the next section we shall demonstrate this option in simulations, showing that it leads to the same reconstruction result as the optimal selection method. Note, however, that our theoretical guarantees do not cover this case.

6.2. ACoSaMP Guarantees

Having the results for AIHT and AHTP we turn to ACoSaMP and ASP. We start with a theorem for ACoSaMP. Its proof is based on the proof for CoSaMP in [6].

Theorem 6.8. *Consider the problem $\mathcal{P}$ and apply ACoSaMP with $a = \frac{2\ell-p}{\ell}$. Let $C_{\hat{\mathcal{S}}} = \max(C_\ell, C_{2\ell-p})$ and suppose that there exists $\gamma > 0$ such that*

$$(1 + C_{\hat{\mathcal{S}}})\left(1 - \left(\frac{C_{\hat{\mathcal{S}}}}{(1+\gamma)^2} - (C_{\hat{\mathcal{S}}} - 1)\sigma_{\mathbf{M}}^2\right)\right) < 1. \tag{53}$$

Then, there exists $\delta_{ACoSaMP}(C_{\hat{\mathcal{S}}}, \sigma_{\mathbf{M}}^2, \gamma) > 0$ such that, whenever $\delta_{4\ell-3p} \leq \delta_{ACoSaMP}(C_{\hat{\mathcal{S}}}, \sigma_{\mathbf{M}}^2, \gamma)$, the t-th iteration of the algorithm satisfies

$$\left\|\mathbf{x} - \hat{\mathbf{x}}^t\right\|_2 \leq \rho_1\rho_2 \left\|\mathbf{x} - \hat{\mathbf{x}}^{t-1}\right\|_2 + (\eta_1 + \rho_1\eta_2) \|\mathbf{e}\|_2 , \tag{54}$$

where

$$\eta_1 \triangleq \frac{\sqrt{\frac{2+C_\ell}{1+C_\ell} + 2\sqrt{C_\ell} + C_\ell}\sqrt{1 + \delta_{3\ell-2p}}}{1 - \delta_{4\ell-3p}},$$

$$\eta_2^2 \triangleq \left(\frac{1 + \delta_{3\ell-2p}}{\gamma(1+\alpha)} + \frac{(1 + \delta_{2\ell-p})C_{2\ell-p}}{\gamma(1+\alpha)(1+\gamma)} + \frac{(C_{2\ell-p} - 1)(1+\gamma)\sigma_{\mathbf{M}}^2}{(1+\alpha)(1+\gamma)\gamma}\right),$$

$$\rho_1^2 \triangleq \frac{1 + 2\delta_{4\ell-3p}\sqrt{C_\ell} + C_\ell}{1 - \delta_{4\ell-3p}^2},$$

$$\rho_2^2 \triangleq 1 - \left(\sqrt{\delta_{4\ell-3p}} - \sqrt{\frac{C_{2\ell-p}}{(1+\gamma)^2}\left(1 - \sqrt{\delta_{2\ell-p}}\right)^2 - (C_{2\ell-p} - 1)(1 + \delta_{2\ell-p})\sigma_{\mathbf{M}}^2}\right)^2$$

and

$$\alpha = \frac{\sqrt{\delta_{4\ell-3p}}}{\sqrt{\frac{C_{2\ell-p}}{(1+\gamma)^2}\left(1 - \sqrt{\delta_{2\ell-p}}\right)^2 - (C_{2\ell-p} - 1)(1 + \delta_{2\ell-p})\sigma_{\mathbf{M}}^2} - \sqrt{\delta_{4\ell-3p}}}.$$

Moreover, $\rho_1^2\rho_2^2 < 1$, i.e., the iterates converges.

The constant γ plays a similar role to the constant η of Theorem 6.5. It gives a tradeoff between satisfying the theorem conditions and the noise amplification. However, as opposed to η, the conditions for the noiseless case are achieved when γ tends to zero. An immediate corollary of the above theorem is the following.

Corollary 6.9. *Consider the problem $\mathcal{P}$ and apply ACoSaMP with $a = \frac{2\ell-p}{\ell}$. If (53) holds and $\delta_{4\ell-3p} < \delta_{ACoSaMP}(C_{\hat{\mathcal{S}}}, \sigma^2_{\mathbf{M}}, \gamma)$, where $C_{\hat{\mathcal{S}}}$ and γ are as in Theorem 6.8 and $\delta_{ACoSaMP}(C_{\hat{\mathcal{S}}}, \sigma^2_{\mathbf{M}}, \gamma)$ is a constant guaranteed to be greater than zero whenever (42) is satisfied, then for any*

$$t \geq t^* = \left\lceil \frac{\log(\|\mathbf{x}\|_2 \,/\, \|\mathbf{e}\|_2)}{\log(1/\rho_1\rho_2)} \right\rceil,$$

$$\left\|\mathbf{x} - \hat{\mathbf{x}}^{t^*}_{ACoSaMP}\right\|_2 \leq \left(1 + \frac{1-(\rho_1\rho_2)^{t^*}}{1-\rho_1\rho_2}\,(\eta_1 + \rho_1\eta_2)\right)\|\mathbf{e}\|_2\,, \tag{55}$$

implying that ACoSaMP leads to a stable recovery. The constants η_1, η_2, ρ_1 and ρ_2 are the same as in Theorem 6.8.

Proof: By using (54) and recursion we have that

$$\begin{aligned}\left\|\mathbf{x} - \hat{\mathbf{x}}^{t^*}_{\text{ACoSaMP}}\right\|_2 &\leq (\rho_1\rho_2)^{t^*}\left\|\mathbf{x} - \hat{\mathbf{x}}^{0}_{\text{ACoSaMP}}\right\|_2 \\ &+(1+\rho_1\rho_2+(\rho_1\rho_2)^2+\ldots(\rho_1\rho_2)^{t^*-1})\,(\eta_1+\rho_1\eta_2)\,\|\mathbf{e}\|_2\,.\end{aligned} \tag{56}$$

Since $\hat{\mathbf{x}}^0_{\text{ACoSaMP}} = 0$, after t^* iterations, one has

$$(\rho_1\rho_2)^{t^*}\left\|\mathbf{x} - \hat{\mathbf{x}}^{0}_{\text{ACoSaMP}}\right\|_2 = (\rho_1\rho_2)^{t^*}\,\|\mathbf{x}\|_2 \leq \|\mathbf{e}\|_2\,. \tag{57}$$

By using the equation of geometric series with (56) and plugging (57) into it, we get (55). □

We turn now to prove the theorem. Instead of presenting the proof directly, we divide the proof into several lemmas. The first lemma gives a bound for $\|\mathbf{x}-\mathbf{w}\|_2$ as a function of $\|\mathbf{e}\|_2$ and $\left\|\mathbf{P}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2$.

Lemma 6.10. *Consider the problem $\mathcal{P}$ and apply ACoSaMP with $a = \frac{2\ell-p}{\ell}$. For each iteration we have*

$$\|\mathbf{x}-\mathbf{w}\|_2 \quad\leq\quad \frac{1}{\sqrt{1-\delta^2_{4\ell-3p}}}\left\|\mathbf{P}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2 + \frac{\sqrt{1+\delta_{3\ell-2p}}}{1-\delta_{4\ell-3p}}\,\|\mathbf{e}\|_2\,. \tag{58}$$

The second lemma bounds $\left\|\mathbf{x} - \hat{\mathbf{x}}^t_{\text{ACoSaMP}}\right\|_2$ in terms of $\left\|\mathbf{P}_{\tilde{\Lambda}^t}(\mathbf{x} - \hat{\mathbf{x}}^t_{\text{ACoSaMP}})\right\|_2$ and $\|\mathbf{e}\|_2$ using the first lemma.

Lemma 6.11. *Consider the problem $\mathcal{P}$ and apply ACoSaMP with $a = \frac{2\ell-p}{\ell}$. For each iteration we have*

$$\left\|\mathbf{x} - \hat{\mathbf{x}}^t\right\|_2 \leq \rho_1\left\|\mathbf{P}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2 + \eta_1\,\|\mathbf{e}\|_2\,, \tag{59}$$

where η_1 and ρ_1 are the same constants as in Theorem 6.8.

The last lemma bounds $\left\|\mathbf{P}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2$ with $\left\|\mathbf{x} - \hat{\mathbf{x}}^{t-1}_{\text{ACoSaMP}}\right\|_2$ and $\|\mathbf{e}\|_2$.

Lemma 6.12. *Consider the problem $\mathcal{P}$ and apply ACoSaMP with $a = \frac{2\ell-p}{\ell}$. if*

$$C_{2\ell-p} < \frac{\sigma^2_{\mathbf{M}}(1+\gamma)^2}{\sigma^2_{\mathbf{M}}(1+\gamma)^2 - 1}, \tag{60}$$

then there exists $\tilde{\delta}_{ACoSaMP}(C_{2\ell-p}, \sigma^2_{\mathbf{M}}, \gamma) > 0$ such that for any $\delta_{2\ell-p} < \tilde{\delta}_{ACoSaMP}(C_{2\ell-p}, \sigma^2_{\mathbf{M}}, \gamma)$

$$\left\|\mathbf{P}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2 \leq \eta_2\,\|\mathbf{e}\|_2 + \rho_2\left\|\mathbf{x} - \hat{\mathbf{x}}^{t-1}\right\|_2. \tag{61}$$

The constants η_2 and ρ_2 are as defined in Theorem 6.8.

The proofs of Lemmas 6.10, 6.11 and 6.12 appear in Appendix D, Appendix E and Appendix F respectively. With the aid of the above three lemmas we turn to prove Theorem 6.8.

Proof:[Proof of Theorem 6.8] Remark that since $1 + C_{\hat{\mathcal{S}}} > 1$ we have that (53) implies $\frac{C_{\hat{\mathcal{S}}}}{(1+\gamma)^2} - (C_{\hat{\mathcal{S}}} - 1)\sigma_{\mathbf{M}}^2 \geq 0$. Because of that the condition in (60) in Lemma 6.12 holds. Substituting the inequality of Lemma 6.12 into the inequality of Lemma 6.11 gives (54). The iterates convergence if $\rho_1^2\rho_2^2 = \frac{1+2\delta_{4\ell-3p}\sqrt{C_\ell}+C_\ell}{1-\delta_{4\ell-3p}^2}\rho_2^2 < 1$. By noticing that $\rho_2^2 < 1$ it is enough to require $\frac{1+C_\ell}{1-\delta_{4\ell-3p}^2}\rho_2^2 + \frac{2\delta_{4\ell-3p}\sqrt{C_\ell}}{1-\delta_{4\ell-3p}^2} < 1$. The last is equivalent to

$$(1 + C_\ell)\left(1 - \left(\sqrt{\delta_{4\ell-3p}} - \sqrt{\frac{C_{2\ell-p}}{(1+\gamma)^2}\left(1 - \sqrt{\delta_{2\ell-p}}\right)^2 - (C_{2\ell-p} - 1)(1 + \delta_{2\ell-p})\sigma_{\mathbf{M}}^2}\right)^2\right) \tag{62}$$
$$+ 2\delta_{4\ell-3p}\sqrt{C_\ell} - 1 + \delta_{4\ell-3p}^2 < 0.$$

It is easy to verify that $\zeta(C, \delta) \triangleq \frac{C}{(1+\gamma)^2}\left(1 - \sqrt{\delta}\right)^2 - (C - 1)(1 + \delta)\sigma_{\mathbf{M}}^2$ is a decreasing function of both δ and C for $0 \leq \delta \leq 1$ and $C > 1$. Since $1 \leq C_{2\ell-p} \leq C_{\hat{\mathcal{S}}}$, $\delta_{2\ell-p} \leq \delta_{4\ell-3p}$ and $\delta \geq 0$ we have that $\zeta(C_{\hat{\mathcal{S}}}, \delta_{4\ell-3p}) \leq \zeta(C_{2\ell-p}, \delta_{4\ell-3p}) \leq \zeta(C_{2\ell-p}, \delta_{2\ell-p}) \leq \zeta(1, 0) = \frac{1}{(1+\gamma)^2} \leq 1$. Thus we have that $-1 \leq -(\sqrt{\delta_{4\ell-3p}} - \zeta(C_{2\ell-p}, \delta_{2\ell-p}))^2 \leq -\delta_{4\ell-3p} + 2\sqrt{\delta_{4\ell-3p}} - \zeta(C_{\hat{\mathcal{S}}}, \delta_{4\ell-3p})$. Combining this with the fact that $C_\ell \leq C_{\hat{\mathcal{S}}}$ provides the following guarantee for $\rho_1^2\rho_2^2 < 1$,

$$(1 + C_{\hat{\mathcal{S}}})\Big(1 - \delta_{4\ell-3p} + 2\sqrt{\delta_{4\ell-3p}} \tag{63}$$
$$-\frac{C_{\hat{\mathcal{S}}}}{(1+\gamma)^2}\left(1 - 2\sqrt{\delta_{4\ell-3p}} + \delta_{4\ell-3p}\right) + (C_{\hat{\mathcal{S}}} - 1)(1 + \delta_{4\ell-3p})\sigma_{\mathbf{M}}^2\Big) + 2\delta_{4\ell-3p}\sqrt{C_{\hat{\mathcal{S}}}} - 1 + \delta_{4\ell-3p}^2 < 0.$$

Let us now assume that $\delta_{4\ell-3p} \leq \frac{1}{2}$. This necessarily means that $\delta_{\text{ACoSaMP}} \leq \frac{1}{2}$ in the end. This assumption implies $\delta_{4\ell-3p}^2 \leq \frac{1}{2}\delta_{4\ell-3p}$. Using this and gathering coefficients, we now consider the condition

$$(1 + C_{\hat{\mathcal{S}}})\left(1 - \frac{C_{\hat{\mathcal{S}}}}{(1+\gamma)^2} + (C_{\hat{\mathcal{S}}} - 1)\sigma_{\mathbf{M}}^2\right) - 1 + 2(1 + C_{\hat{\mathcal{S}}})\left(1 + \frac{C_{\hat{\mathcal{S}}}}{(1+\gamma)^2}\right)\sqrt{\delta_{4\ell-3p}} \tag{64}$$
$$+\left((1 + C_{\hat{\mathcal{S}}})\left(-1 - \frac{C_{\hat{\mathcal{S}}}}{(1+\gamma)^2} + (C_{\hat{\mathcal{S}}} - 1)\sigma_{\mathbf{M}}^2\right) + 2\sqrt{C_{\hat{\mathcal{S}}}} + \frac{1}{2}\right)\delta_{4\ell-3p} < 0.$$

The expression on the LHS is a quadratic function of $\sqrt{\delta_{4\ell-3p}}$. Note that since (53) holds the constant term in the quadratic function is negative. This guarantees the existence of a range of values $[0, \delta_{\text{ACoSaMP}}(C_{\hat{\mathcal{S}}}, \sigma_{\mathbf{M}}^2, \gamma)]$ for $\delta_{4\ell-3p}$ for which (64) holds, where $\delta_{\text{ACoSaMP}}(C_{\hat{\mathcal{S}}}, \sigma_{\mathbf{M}}^2, \gamma)$ is the square of the positive solution of the quadratic function. In case of two positive solutions we should take the smallest among them – in this case the coefficient of $\delta_{4\ell-3p}$ in (64) will be positive.

Looking back at the proof of the theorem, we observe that the value of the constant $\delta_{\text{ACoSaMP}}(C_{\hat{\mathcal{S}}}, \sigma_{\mathbf{M}}^2, \gamma)$ can potentially be improved: at the beginning of the proof, we have used $\rho_2^2 \leq 1$. By the end, we obtained $\rho_2^2 \leq \rho_1^{-2} \leq 0.25$ since $\rho_1 > 2$. If we were to use this bound at the beginning, we would have obtained better constant $\delta_{\text{ACoSaMP}}(C_{\hat{\mathcal{S}}}, \sigma_{\mathbf{M}}^2, \gamma)$. □

6.3. ASP Guarantees

Having the result of ACoSaMP we turn to derive a similar result for ASP. The technique for deriving a result for ASP based on the result of ACoSaMP is similar to the one we used to derive a result for AHTP from the result of AIHT.

Theorem 6.13. *Consider the problem $\mathcal{P}$ and apply ASP with $a = \frac{2\ell-p}{\ell}$. If* (53) *holds and $\delta_{4\ell-3p} \leq \delta_{ASP}(C_{\hat{\mathcal{S}}}, \sigma_{\mathbf{M}}^2, \gamma)$, where $C_{\hat{\mathcal{S}}}$ and γ are as in Theorem 6.8, and $\delta_{ASP}(C_{\hat{\mathcal{S}}}, \sigma_{\mathbf{M}}^2, \gamma)$ is a constant guaranteed to be greater than zero whenever* (53) *is satisfied, then the t-th iteration of the algorithm satisfies*

$$\left\|\mathbf{x} - \hat{\mathbf{x}}_{ASP}^t\right\|_2 \leq \frac{1 + \delta_{2\ell-p}}{1 - \delta_{2\ell-p}}\rho_1\rho_2\left\|\mathbf{x} - \hat{\mathbf{x}}_{ASP}^{t-1}\right\|_2 + \left(\frac{1 + \delta_{2\ell-p}}{1 - \delta_{2\ell-p}}(\eta_1 + \rho_1\eta_2) + \frac{2}{1 - \delta_{2\ell-p}}\right)\|\mathbf{e}\|_2 . \tag{65}$$

and the iterates converges, i.e., $\rho_1^2\rho_2^2 < 1$. The constants η_1, η_2, ρ_1 and ρ_2 are the same as in Theorem 6.8.

Proof: We first note that according to the selection rule of $\hat{\mathbf{x}}_{\text{ASP}}$ we have that

$$\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}_{\text{ASP}}^t\right\|_2 \leq \left\|\mathbf{y} - \mathbf{M}\mathbf{Q}_{\hat{\Lambda}^t}\mathbf{w}\right\|_2 . \tag{66}$$

Using the triangle inequality and the fact that $\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{e}$ for both the LHS and the RHS we have

$$\left\|\mathbf{M}(\mathbf{x} - \hat{\mathbf{x}}_{\text{ASP}}^t)\right\|_2 - \|\mathbf{e}\|_2 \leq \left\|\mathbf{M}(\mathbf{x} - \mathbf{Q}_{\hat{\Lambda}^t}\mathbf{w})\right\|_2 + \|\mathbf{e}\|_2 .$$

Using the $\mathbf{\Omega}$-RIP property of $\mathbf{M}$ with the fact that $\mathbf{x}$, $\hat{\mathbf{x}}_{\text{ASP}}$ and $\mathbf{Q}_{\hat{\Lambda}^t}\mathbf{w}$ are ℓ-cosparse we have

$$\left\|\mathbf{x} - \hat{\mathbf{x}}_{\text{ASP}}^t\right\|_2 \leq \frac{1+\delta_{2\ell-p}}{1-\delta_{2\ell-p}} \left\|\mathbf{x} - \mathbf{Q}_{\hat{\Lambda}^t}\mathbf{w}\right\|_2 + \frac{2}{1-\delta_{2\ell-p}} \|\mathbf{e}\|_2 .$$

Noticing that $\mathbf{Q}_{\hat{\Lambda}^t}\mathbf{w}$ is the solution we get in one iteration of ACoSaMP with initialization of $\hat{\mathbf{x}}_{\text{ASP}}^{t-1}$, we can combine the above with the result of Theorem 6.8 getting (65). For $\frac{1+\delta_{2\ell-p}}{1-\delta_{2\ell-p}}\rho_1\rho_2 < 1$ to hold we need that

$$\frac{1 + 2\delta_{4\ell-3p}\sqrt{C_\ell} + C_\ell}{(1-\delta_{4\ell-3p})^2}\left(1 - \left(\left(\frac{\sqrt{\tilde{C}_{2\ell-p}}}{1+\gamma} + 1\right)\sqrt{\delta_{4\ell-3p}} - \frac{\sqrt{\tilde{C}_{2\ell-p}}}{1+\gamma}\right)^2\right) < 1. \tag{67}$$

Remark that the above differs from what we have for ACoSaMP only in the denominator of the first element in the LHS. In ACoSaMP $1 - \delta_{4\ell-3p}^2$ appears instead of $(1-\delta_{4\ell-3p})^2$. Thus, Using a similar process to the one in the proof of ACoSaMP we can show that (67) holds if the following holds

$$\begin{aligned}&(1+C_{\hat{\mathcal{S}}})\left(1 - \frac{C_{\hat{\mathcal{S}}}}{(1+\gamma)^2} + (C_{\hat{\mathcal{S}}} - 1)\sigma_{\mathbf{M}}^2\right) - 1 + 2(1+C_{\hat{\mathcal{S}}})\left(1 + \frac{C_{\hat{\mathcal{S}}}}{(1+\gamma)^2}\right)\sqrt{\delta_{4\ell-3p}} \\ &+ \left((1+C_{\hat{\mathcal{S}}})\left(-1 - \frac{C_{\hat{\mathcal{S}}}}{(1+\gamma)^2} + (C_{\hat{\mathcal{S}}} - 1)\sigma_{\mathbf{M}}^2\right) + 2\sqrt{C_{\hat{\mathcal{S}}}} + 2\right)\delta_{4\ell-3p} < 0.\end{aligned} \tag{68}$$

Notice that the only difference of the above compared to (64) is that we have +2 instead of +0.5 in the coefficient of $\delta_{4\ell-3p}$ and this is due to the difference we mentioned before in the denominator in (67). The LHS of (68) is a quadratic function of $\sqrt{\delta_{4\ell-3p}}$. As before, we notice that if (53) holds then the constant term of the above is positive and thus $\delta_{\text{ASP}}(C_{\hat{\mathcal{S}}}, \sigma_{\mathbf{M}}^2, \gamma) \geq 0$ exists and is the square of the positive solution of the quadratic function. □

Having Theorem 6.13 we can immediately have the following corollary which is similar to the one we have for ACoSaMP. The proof resembles the one of Corollary 6.9 and omitted.

Corollary 6.14. *Consider the problem $\mathcal{P}$ and apply ASP with $a = \frac{2\ell-p}{\ell}$. If* (53) *holds and $\delta_{4\ell-3p} \leq \delta_{ASP}(C_{\hat{\mathcal{S}}}, \sigma_{\mathbf{M}}^2, \gamma)$, where $C_{\hat{\mathcal{S}}}$ and γ are as in Theorem 6.8, and $\delta_{ASP}(C_{\hat{\mathcal{S}}}, \sigma_{\mathbf{M}}^2, \gamma)$ is a constant guaranteed to be greater than zero whenever* (42) *is satisfied, then for any*

$$t \geq t^* = \left\lfloor \frac{\log(\|\mathbf{x}\|_2 \,/\, \|\mathbf{e}\|_2)}{\log(1/\frac{1+\delta_{2\ell-p}}{1-\delta_{2\ell-p}}\rho_1\rho_2)} \right\rfloor,$$

$$\left\|\mathbf{x}_{ASP}^t - \mathbf{x}\right\|_2 \leq \left(1 + \frac{1 - \left(\frac{1+\delta_{2\ell-p}}{1-\delta_{2\ell-p}}\rho_1\rho_2\right)^t}{1 - \frac{1+\delta_{2\ell-p}}{1-\delta_{2\ell-p}}\rho_1\rho_2} \cdot \left(\frac{1+\delta_{2\ell-p}}{1-\delta_{2\ell-p}}(\eta_1 + \rho_1\eta_2) + \frac{2}{1-\delta_{2\ell-p}}\right)\right)\|\mathbf{e}\|_2 . \tag{69}$$

implying that ASP leads to a stable recovery. The constants η_1, η_2, ρ_1 and ρ_2 are the same as in Theorem 6.8.

6.4. Non-Exact Cosparse Case

In the above guarantees we have assumed that the signal $\mathbf{x}$ is ℓ-cosparse. In many cases, it is not exactly ℓ-cosparse but only nearly so. Denote by $\mathbf{x}^\ell = \mathbf{Q}_{\mathcal{S}^*_\ell(\mathbf{x})}\mathbf{x}$ the best ℓ-cosparse approximation of $\mathbf{x}$, we have the following theorem that provides us with a guarantee also for this case. Similar result exists also in the synthesis case for the synthesis-ℓ_1 minimization problem [39].

Theorem 6.15. *Consider a variation of problem $\mathcal{P}$ where $\mathbf{x}$ is a general vector, and apply either AIHT or AHTP both with either constant or changing step size; or ACoSaMP or ASP with $a = \frac{2\ell-p}{\ell}$, and all are used with a zero initialization. Under the same conditions of Theorems 6.2 and 6.3 we have for any $t \geq t^*$*

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq \left\|\mathbf{x} - \mathbf{x}^\ell\right\|_2 + c\left\|\mathbf{M}(\mathbf{x} - \mathbf{x}^\ell)\right\|_2 + c\|\mathbf{e}\|_2, \tag{70}$$

where t^* and c are the constants from Theorems 6.2 and 6.3.

Proof: First we notice that we can rewrite $\mathbf{y} = \mathbf{M}\mathbf{x}^\ell + \mathbf{M}(\mathbf{x} - \mathbf{x}^\ell) + \mathbf{e}$. Denoting $\mathbf{e}^\ell = \mathbf{M}(\mathbf{x} - \mathbf{x}^\ell) + \mathbf{e}$ we can use Theorems 6.2 and 6.3 to recover $\mathbf{x}^\ell$ and have

$$\left\|\mathbf{x}^\ell - \hat{\mathbf{x}}\right\|_2 \leq c\left\|\mathbf{e}^\ell\right\|_2. \tag{71}$$

Using the triangle inequality for $\|\mathbf{x} - \hat{\mathbf{x}}\|_2$ with the above gives

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq \left\|\mathbf{x} - \mathbf{x}^\ell\right\|_2 + \left\|\mathbf{x}^\ell - \hat{\mathbf{x}}\right\|_2 \leq \left\|\mathbf{x} - \mathbf{x}^\ell\right\|_2 + c\left\|\mathbf{e}^\ell\right\|_2. \tag{72}$$

Using again the triangle inequality for $\left\|\mathbf{e}^\ell\right\|_2 \leq \|\mathbf{e}\|_2 + \left\|\mathbf{M}(\mathbf{x} - \mathbf{x}^\ell)\right\|_2$ provides us with the desired result. □

6.5. Theorem Conditions

Having the results of the theorems we ask ourselves whether their conditions are feasible. As we have seen in Section 3, the requirement on the $\mathbf{\Omega}$-RIP for many non-trivial matrices. In addition, as we have seen in the introduction of this section we need C_ℓ and $C_{2\ell-p}$ to be one or close to one for satisfying the conditions of the theorems. Using the thresholding in (25) for cosupport selection with a unitary $\mathbf{\Omega}$ satisfies the conditions in a trivial way since $C_\ell = C_{2\ell-p} = 1$. This case coincides with the synthesis model for which we already have theoretical guarantees. As shown in Section 4, optimal projection schemes exist for $\mathbf{\Omega}_{\text{1D-DIF}}$ and $\mathbf{\Omega}_{\text{FUS}}$ which do not belong to the the synthesis framework. For a general $\mathbf{\Omega}$, a general projection scheme is not known and if the thresholding method is used the constants in (25) do not equal one and are not even expected to be close to one [27]. It is interesting to ask whether there exists an efficient general projection scheme that guarantees small constants for any given operator $\mathbf{\Omega}$, or for specifically structured $\mathbf{\Omega}$. We leave these questions as subject for future work. Instead, we show empirically in the next section that a weaker projection scheme that does not fulfill all the requirements of the theorems leads to a good reconstruction result. This suggests that even in the absence of good near optimal projections we may still use the algorithms practically.

6.6. Comparison to Other Works

Among the existing theoretical works that studied the performance of analysis algorithms [18, 22, 26], the result that resembles ours is the result for ℓ_1-analysis in [21]. This work analyzed the ℓ_1-analysis minimization problem with a synthesis perspective. The analysis dictionary $\mathbf{\Omega}$ was replaced with the conjugate of a synthesis dictionary $\mathbf{D}$ which is assumed to be a tight frame, resulting with the following minimization problem.

$$\hat{\mathbf{x}}_{A-\ell_1} = \underset{\mathbf{z}}{\operatorname{argmin}} \|\mathbf{D}^*\mathbf{z}\|_1 \quad s.t. \quad \|\mathbf{y} - \mathbf{M}\mathbf{z}\|_2 \leq \epsilon. \tag{73}$$

It was shown that if $\mathbf{M}$ has the $\mathbf{D}$-RIP [21, 29] with $\delta_{7k} < 0.6$, an extension of the synthesis RIP, then

$$\left\|\hat{\mathbf{x}}_{A-\ell_1} - \mathbf{x}\right\|_2 \leq \tilde{C}_{\ell_1}\epsilon + \frac{\|\mathbf{D}^*\mathbf{x} - [\mathbf{D}^*\mathbf{x}]_k\|_1}{\sqrt{k}}. \tag{74}$$

We say that a matrix $\mathbf{M}$ has a $\mathbf{D}$-RIP with a constant δ_k if for any signal $\mathbf{z}$ that has a k-sparse representation under $\mathbf{D}$

$$(1-\delta_k)\left\|\mathbf{z}\right\|_2^2 \leq \left\|\mathbf{M}\mathbf{z}\right\|_2^2 \leq (1+\delta_k)\left\|\mathbf{z}\right\|_2^2. \tag{75}$$

The authors in [21] presented this result as a synthesis result that allows linear dependencies in $\mathbf{D}$ at the cost of limiting the family of signals to be those for which $\|\mathbf{D}^*\mathbf{x} - [\mathbf{D}^*\mathbf{x}]_k\|_1$ is small. However, having the analysis perspective, we can realize that they provided a recovery guarantee for ℓ_1-analysis under the new analysis model for the case that $\mathbf{\Omega}$ is a tight frame. An easy way to see it is to observe that for an ℓ-cosparse signal $\mathbf{x}$, setting $k = p - \ell$, we have that $\left\|\mathbf{\Omega}\mathbf{x} - [\mathbf{\Omega}^*\mathbf{x}]_{p-\ell}\right\|_1 = 0$ and thus in the case $\epsilon = 0$ we get that (74) guarantees the recovery of $\mathbf{x}$ by using (73) with $\mathbf{D}^* = \mathbf{\Omega}$. Thus, though the result in [21] was presented as a reconstruction guarantee for the synthesis model, it is actually a guarantee for the analysis model.

Our main difference from [21] is that the proof technique relies on the analysis model and not on the synthesis one and that the results presented here are for general operators and not only for tight frames. For instance, the operators $\mathbf{\Omega}_{\text{1D-DIF}}$ and $\mathbf{\Omega}_{\text{FUS}}$ for which the guarantees hold are not tight frames where $\mathbf{\Omega}_{\text{1D-DIF}}$ is not even a frame. However, the drawback of our approached compared to the work in [21] is that it is still not known how to perform an optimal or a near optimal projection for a tight frame.

In the non-exact sparse case our results differ from the one in (74) in the sense that it looks at the projection error and not at the values of $\mathbf{\Omega}\mathbf{x}$. It would be interesting to see if there is a connection between the two and whether one implies the other.

A recent work has studied the ℓ_1-analysis minimization with the 2D-DIF operator, also known as anisotropic two dimensional total-variation (2D-TV) [40]. It would be interesting to see whether similar results can be achieved for the greedy-like techniques proposed here with 2D-DIF.

7. Experiments

In this section we repeat some of the experiments performed in [18] for the noiseless case ($\mathbf{e} = 0$) and some of the experiments performed in [23] for the noisy case[4].

7.1. Targeted Cosparsity

Just as in the synthesis counterpart of the proposed algorithms, where a target sparsity level k must be selected before running the algorithms, we have to choose the targeted cosparsity level which will dictate the projection steps. In the synthesis case it is known that it may be beneficial to over-estimate the sparsity k. Similarly in the analysis framework the question arises: In terms of recovery performance, does it help to under-estimate the cosparsity ℓ? A tentative positive answer comes from the following heuristic: Let $\tilde{\Lambda}$ be a subset of the cosupport Λ of the signal $\mathbf{x}$ with $\tilde{\ell} := |\tilde{\Lambda}| < \ell = |\Lambda|$. According to Proposition 3 in [18]

$$\kappa_{\mathbf{\Omega}}(\tilde{\ell}) \leq \frac{m}{2} \tag{76}$$

is a sufficient condition to identify $\tilde{\Lambda}$ in order to recover $\mathbf{x}$ from the relations $\mathbf{y} = \mathbf{M}\mathbf{x}$ and $\mathbf{\Omega}_{\tilde{\Lambda}}\mathbf{x} = 0$. $\kappa_{\mathbf{\Omega}}(\tilde{\ell}) = \max_{\tilde{\Lambda}\in\mathcal{L}_{\tilde{\ell}}} \dim(\mathcal{W}_{\tilde{\Lambda}})$ is a function of $\tilde{\ell}$. Therefore, we can replace ℓ with the smallest $\tilde{\ell}$ that satisfies (76) as the effective cosparsity in the algorithms. Since it is easier to identify a smaller cosupport set it is better to run the algorithm with the smallest possible value of $\tilde{\ell}$, in the absence of noise. In the presence of noise, larger values of ℓ allows a better denoising. Note, that in some cases the smallest possible value of $\tilde{\ell}$ will be larger than the actual cosparsity of $\mathbf{x}$. In this case we cannot replace ℓ with $\tilde{\ell}$.

We take two examples for selecting $\tilde{\ell}$. The first is for $\mathbf{\Omega}$ which is in general position and the second is for $\mathbf{\Omega}_{2D-DIF}$, the finite difference analysis operator that computes horizontal and vertical discrete derivatives of an image which is strongly connected to the total variation (TV) norm minimization as noted before. For $\mathbf{\Omega}$ that is in general position $\kappa_{\mathbf{\Omega}}(\tilde{\ell}) = \max(d - \ell, 0)$ [18]. In this case we choose

$$\tilde{\ell} = \min\left(d - \frac{m}{2}, \ell\right). \tag{77}$$

[4] A matlab package with code for the experiments performed in this paper is in preparation for an open source distribution.

For $\mathbf{\Omega}_{DIF}$ we have $\kappa_{\mathbf{\Omega}_{DIF}}(\tilde{\ell}) \geq d - \frac{\ell}{2} - \sqrt{\frac{\ell}{2}} - 1$ [18] and

$$\tilde{\ell} = \lceil \min((-1/\sqrt{2} + \sqrt{2d - m - 1.5})^2, \ell) \rceil. \tag{78}$$

Replacing ℓ with $\tilde{\ell}$ is more relevant to AIHT and AHTP than ACoSaMP and ASP since in the last we intersect cosupport sets and thus the estimated cosupport set need to be large enough to avoid empty intersections. Thus, for $\mathbf{\Omega}$ in general position we use the true cosparsity level for ACoSaMP and ASP. For $\mathbf{\Omega}_{DIF}$, where linear dependencies occur, the corank does not equal the cosparsity and we use $\tilde{\ell}$ instead of ℓ since it will be favorable to run the algorithm targeting a cosparsity level in the middle. In this case ℓ tends to be very large and it is more likely to have non-empty intersections .

7.2. Phase Diagrams for Synthetic Signals in the Noiseless Case

We begin with with synthetic signals in the noiseless case. We test the performance of AIHT with a constant step-size, AIHT with an adaptive changing step-size, AHTP with a constant step-size, AHTP with an adaptive changing step-size, ACoSaMP with $a = \frac{2\ell - p}{\ell}$, ACoSaMP with $a = 1$, ASP with $a = \frac{2\ell - p}{\ell}$ and ASP with $a = 1$. We compare the results to those of A-ℓ_1-minimization [20] and GAP [18]. We use a random matrix $\mathbf{M}$ and a random tight frame with $d = 120$ and $p = 144$, where each entry in the matrices is drawn independently from the Gaussian distribution.

We draw a phase transition diagram [41] for each of the algorithms. We test 20 different possible values of m and 20 different values of ℓ and for each pair repeat the experiment 50 times. In each experiment we check whether we have a perfect reconstruction. White cells in the diagram denotes a perfect reconstruction in all the experiments of the pair and black cells denotes total failure in the reconstruction. The values of m and ℓ are selected according to the following formula:

$$m = \delta d \qquad \ell = d - \rho m, \tag{79}$$

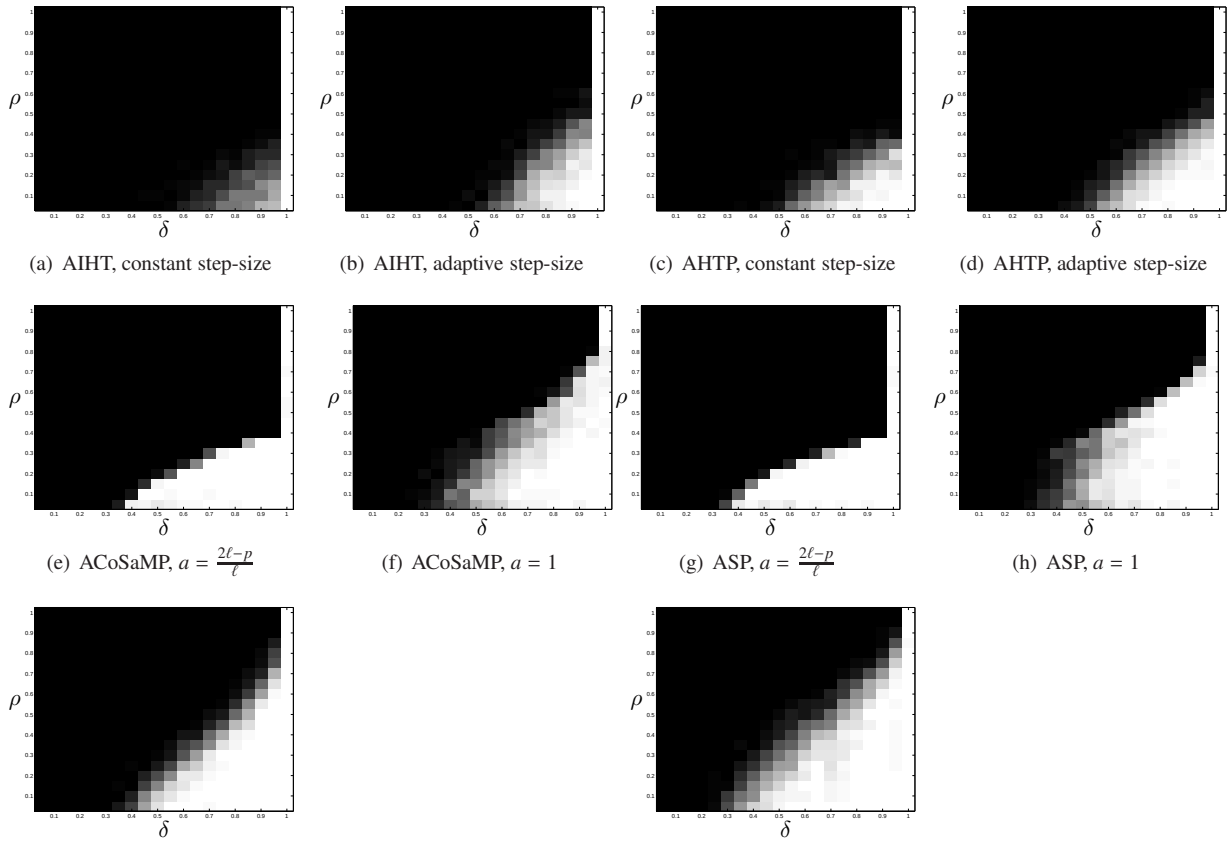
where δ, the sampling rate, is the x-axis of the phase diagram and ρ, the ratio between the cosparsity of the signal and the number of measurements, is the y-axis.

Figure 2 presents the reconstruction results of the algorithms. It should be observed that AIHT and AHTP have better performance using the adaptive step-size than using the constant step-size. The optimal step-size has similar reconstruction result like the adaptive one and thus not presented. For ACoSaMP and ASP we observe that it is better to use $a = 1$ instead of $a = \frac{2\ell - p}{\ell}$. Compared to each other we see that ACoSaMP and ASP achieve better recovery than AHTP and AIHT. Between the last two, AHTP is better. Though AIHT has inferior behavior, we should mention that with regards to running time AIHT is the most efficient. Afterwards we have AHTP and then ACoSaMP and ASP. Compared to ℓ_1 and GAP we observe that ACoSaMP and ASP have competitive results.

With the above observations, we turn to test operators with higher redundancy and see the effect of linear dependencies in them. We test two operators. The first is a random tight frame as before but with redundancy factor of 2. The second is the two dimensional finite difference operator $\mathbf{\Omega}_{\text{2D-DIF}}$. In Fig. 3 we present the phase diagrams for both operators using AIHT with an adaptive changing step-size, AHTP with an adaptive changing step-size, ACoSaMP with $a = 1$, and ASP with $a = 1$. As observed before, also in this case the ACoSaMP and ASP outperform AIHT and AHTP in both cases and AHTP outperform AIHT. We mention again that the better performance comes at the cost of higher complexity. In addition, as we expected, having redundancies in $\mathbf{\Omega}$ results with a better recovery.

7.3. Reconstruction of High Dimensional Images in the Noisy Case

We turn now to test the methods for high dimensional signals. We use RASP and RACoSaMP (relaxed versions of ASP and ACoSaMP defined in Section 5.4) for the reconstruction of the *Shepp-Logan phantom* from few number of measurements. The sampling operator is a two dimensional Fourier transform that measures only a certain number of radial lines from the Fourier transform. The cosparse operator is $\mathbf{\Omega}_{\text{2D-DIF}}$ and the cosparsity used is the actual cosparsity of the signal under this operator ($\ell = 128014$). The phantom image is presented in Fig. 4(a). Using the RACoSaMP and RASP we get a perfect reconstruction using only 15 radial lines, i.e., only $m = 3782$ measurements out of $d = 65536$ which is less then 6 percent of the data in the original image. The algorithms requires less than 20 iterations for having this perfect recovery. For AIHT and RAHTP we achieve a reconstruction which is only close to



(a) AIHT, constant step-size (b) AIHT, adaptive step-size (c) AHTP, constant step-size (d) AHTP, adaptive step-size

(e) ACoSaMP, $a = \frac{2\ell-p}{\ell}$ (f) ACoSaMP, $a = 1$ (g) ASP, $a = \frac{2\ell-p}{\ell}$ (h) ASP, $a = 1$

(i) A-ℓ_1-minimization (j) GAP

Figure 2: Recovery rate for a random tight frame with $p = 144$ and $d = 120$. From left to right, up to bottom: AIHT with a constant step-size, AIHT with an adaptive changing step-size, AHTP with a constant step-size, AHTP with an adaptive changing step-size, ACoSaMP with $a = \frac{2\ell-p}{\ell}$, ACoSaMP with $a = 1$, ASP with $a = \frac{2\ell-p}{\ell}$, ASP with $a = 1$, A-ℓ_1-minimization and GAP.

the original image using 35 radial lines. The reconstruction result of AIHT is presented in Fig 4(b). The advantage of the AIHT, though it has an inferior performance, over the other methods is its running time. While the others need several minutes for each reconstruction, for the AIHT it takes only few seconds to achieve a visually reasonable result.

Exploring the noisy case, we perform a reconstruction using RASP of a noisy measurement of the phantom with 22 radial lines and signal to noise ratio (SNR) of 20. Figure 4(c) presents the noisy image, the result of applying inverse Fourier transform on the measurements, and Fig. 4(d) presents its reconstruction result. Note that for the minimization process we solve conjugate gradients, in each iteration and take only the real part of the result and crop the values of the resulted image to be in the range of $[0, 1]$. We get a peak SNR (PSNR) of $36dB$. We get similar results using RACoSaMP but using more radial lines (25).

8. Discussion and Conclusion

In this work we presented new pursuits for the cosparse analysis model. A theoretical study of these algorithms was performed giving guarantees for stable recovery under the assumptions of the $\mathbf{\Omega}$-RIP and the existence of an optimal or a near optimal projection. We showed that optimal projections exists for some non-trivial operators, i.e., operators that do not take us back to the synthesis case. In addition, we showed experimentally that using simpler kind of projections is possible in order to get good reconstruction results. We demonstrated both in the theoretical and the empirical results that linear dependencies within the analysis dictionary are favorable and enhance the recovery

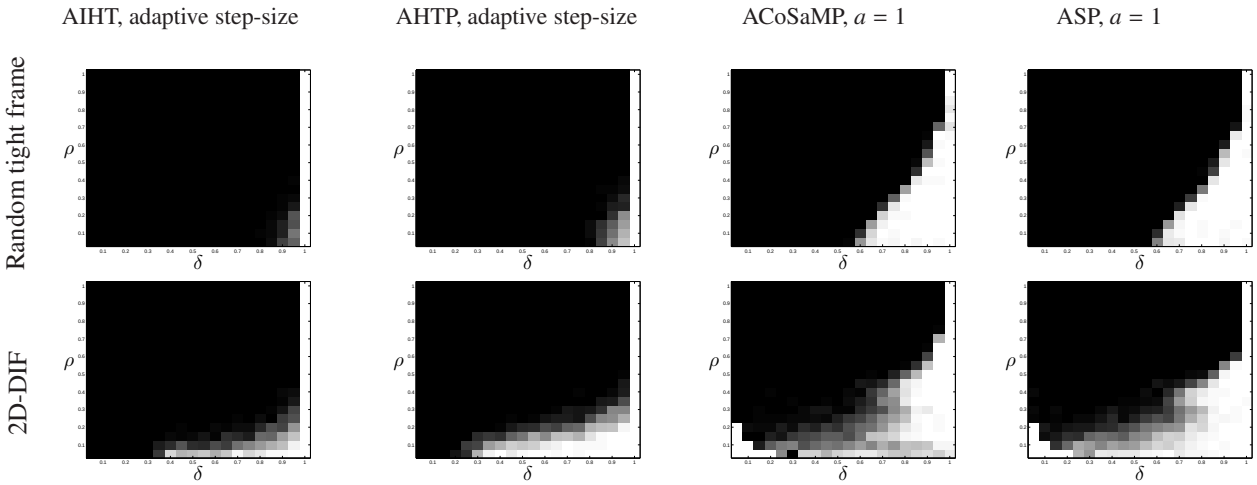
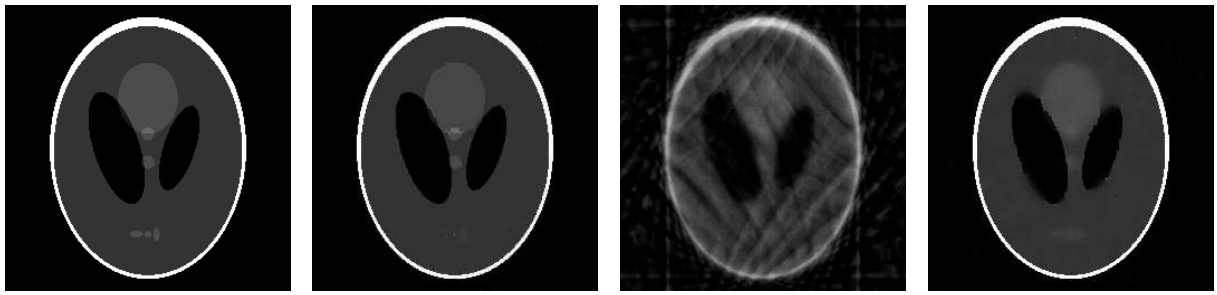


Figure 3: Recovery rate for a random tight frame with $p = 240$ and $d = 120$ (up) and a finite difference operator (bottom). From left to right: AIHT and AHTP with an adaptive changing step-size, and ACoSaMP and ASP with $a = 1$.



(a) Phantom (b) AIHT - noiseless (c) Noisy Phantom (d) RASP - noisy

Figure 4: From left to right: Shepp Logan phantom image, AIHT reconstruction using 35 radial lines, noisy image with SNR of 20 and recovered image using RASP and only 22 radial lines. Note that for the noiseless case RASP and RACoSaMP get a perfect reconstruction using only 15 radial lines.

performance.

We are aware that there are still some open questions in this work and we leave them for future research. This should deal with following:

- Our work assumed the existence of a procedure that finds a cosupport that implies a near optimal projection with a constant C_ℓ. Two examples for optimal cosupport slection schemes were given. However, the existence of an optimal or a near optimal scheme for a general operator is still an open question. The question is: for which types of $\mathbf{\Omega}$ and values of C_ℓ we can find an efficient procedure that implies a near optimal projection.

- As we have seen in the simulations, the thresholding procedure, though not near optimal with the theorems required constants, provides good reconstruction results. A theoretical study of the analysis greedy-like techniques with this cosupport selection scheme is required.

- A family of analysis dictionaries that deserves a special attention is the family of tight frame operators. In synthesis, there is a parallel between the guarantees of ℓ_1-synthesis and the greedy like algorithms. The fact that a guarantee with a tight frame $\mathbf{\Omega}$ exists for ℓ_1-analysis encourage us to believe that similar guarantees exist also for the analysis greedy-like techniques.

- In this paper, the noise $\mathbf{e}$ was considered to be adversarial. Random white Gaussian case was considered for

the synthesis case in [15] resulting with near-oracle performance guarantees. It would be interesting to verify whether this is also the case for the analysis framework.

Appendix A. Proofs of Theorem 3.7 and Theorem 3.8

Theorem 3.7 (Theorem 3.3 in [29]): Let $\mathbf{M} \in \mathbb{R}^{m\times d}$ be a random matrix that satisfies that for any $\mathbf{z} \in \mathbb{R}^d$ and $0 < \tilde{\epsilon} \leq \frac{1}{3}$

$$P\left(\left|\|\mathbf{M}\mathbf{z}\|_2^2 - \|\mathbf{z}\|_2^2\right| \geq \tilde{\epsilon}\,\|\mathbf{z}\|_2^2\right) \leq e^{-\frac{C_{\mathbf{M}} m\tilde{\epsilon}}{2}},$$

where $C_{\mathbf{M}} > 0$ is a constant. For any value of $\epsilon_r > 0$, if

$$m \geq \frac{32}{C_M \epsilon_r^2}\left(\log(\left|\mathcal{L}_r^{\text{corank}}\right|) + (d-r)\log(9/\epsilon_r) + t\right),$$

then $\delta_r^{\text{corank}} \leq \epsilon_r$ with probability exceeding $1 - e^{-t}$.

Theorem 3.8: Under the same setup of Theorem 3.7, for any $\epsilon_\ell > 0$ if

$$m \geq \frac{32}{C_M \epsilon_\ell^2}\left((p-\ell)\log\left(\frac{9p}{(p-\ell)\epsilon_\ell}\right) + t\right),$$

then $\delta_\ell \leq \epsilon_\ell$ with probability exceeding $1 - e^{-t}$.

Proof: Let $\tilde{\epsilon} = \epsilon_r/4$, $B^{d-r} = \{\mathbf{z} \in \mathbb{R}^{d-r}, \|\mathbf{z}\|_2 \leq 1\}$ and Ψ an $\tilde{\epsilon}$-net for B^{d-r} with size $|\Psi| \leq \left(1 + \frac{2}{\tilde{\epsilon}}\right)^{d-r}$ [30]. For any subspace $\mathcal{W}_\Lambda^B = \mathcal{W}_\Lambda \cap B^{d-r}$ such that $\Lambda \in \mathcal{L}_r^{\text{corank}}$ we can build an orthogonal matrix $\mathbf{U}_{\mathbf{\Omega}} \in \mathbb{R}^{d\times(d-r)}$ such that $\mathcal{W}_\Lambda^B = \{\mathbf{U}_\Lambda \mathbf{z}, \mathbf{z} \in \mathbb{R}^{d-r}, \|\mathbf{z}\|_2 \leq 1\} = \mathbf{U}_\Lambda B^{d-r}$. It is easy to see that $\Psi_\Lambda = \mathbf{U}_\Lambda \Psi^{d-r}$ is an $\tilde{\epsilon}$-net for W_Λ^B and that $\Psi_{\mathcal{A}_r^{\text{corank}}} = \cup_{\Lambda\in\mathcal{L}_r^{\text{corank}}}\Psi_\Lambda$ is an $\tilde{\epsilon}$-net for $\mathcal{A}_r^{\text{corank}} \cap B^d$, where $\left|\Psi_{\mathcal{A}_r^{\text{corank}}}\right| \leq \left|\mathcal{L}_r^{\text{corank}}\right|(1 + \frac{2}{\tilde{\epsilon}})^{d-r}$.

We could stop here and use directly Theorem 2.1 from [30] to get the desired result for Theorem 3.7. However, we present the remaining of the proof using a proof technique from [32, 8]. Using union bound and the properties of $\mathbf{M}$ we have that with probability exceeding $1 - \left|\mathcal{L}_r^{\text{corank}}\right|(1 + \frac{2}{\tilde{\epsilon}})^{d-r} e^{-\frac{C_{\mathbf{M}} m\tilde{\epsilon}^2}{2}}$ every $\mathbf{v} \in \Psi_{\mathcal{A}_r^{\text{corank}}}$ satisfies

$$(1-\tilde{\epsilon})\,\|\mathbf{v}\|_2^2 \leq \|\mathbf{M}\mathbf{v}\|_2^2 \leq (1+\tilde{\epsilon})\,\|\mathbf{v}\|_2^2\,. \tag{A.1}$$

According to the definition of δ_r^{corank} it holds that $\sqrt{1+\delta_r^{\text{corank}}} = \sup_{\mathbf{v}\in\mathcal{A}_r^{\text{corank}}\cap B^d}\|\mathbf{M}\mathbf{v}\|_2$. Since $\mathcal{A}_r^{\text{corank}} \cap B^d$ is a compact set there exists $\mathbf{v}_0 \in \mathcal{A}_r^{\text{corank}} \cap B^d$ that achieves the supremum. Denoting by $\tilde{\mathbf{v}}$ its closest vector in $\Psi_{\mathcal{A}_r^{\text{corank}}}$ and using the definition of $\Psi_{\mathcal{A}_r^{\text{corank}}}$ we have $\|\mathbf{v}_0 - \tilde{\mathbf{v}}\|_2 \leq \tilde{\epsilon}$. This yields

$$\begin{aligned}\sqrt{1+\delta_r^{\text{corank}}} = \|\mathbf{M}\mathbf{v}_0\|_2 &\leq \|\mathbf{M}\tilde{\mathbf{v}}\|_2 + \|\mathbf{M}(\mathbf{v}_0 - \tilde{\mathbf{v}})\|_2 \\ &\leq \sqrt{1+\tilde{\epsilon}} + \left\|\mathbf{M}\frac{\mathbf{v}_0 - \tilde{\mathbf{v}}}{\|\mathbf{v}_0 - \tilde{\mathbf{v}}\|_2}\right\|_2 \|\mathbf{v}_0 - \tilde{\mathbf{v}}\|_2 \leq \sqrt{1+\tilde{\epsilon}} + \sqrt{1+\delta_r^{\text{corank}}}\tilde{\epsilon}.\end{aligned} \tag{A.2}$$

The first inequality is due to the triangle inequality; the second one follows from (A.1) and arithmetics; and the last inequality follows from the definition of δ_r^{corank}, the properties of $\tilde{\epsilon}$-net and the fact that $\left\|\frac{\mathbf{v}_0-\tilde{\mathbf{v}}}{\|\mathbf{v}_0-\tilde{\mathbf{v}}\|_2}\right\|_2 = 1$. Reordering (A.2) gives

$$1 + \delta_r^{\text{corank}} \leq \frac{1+\tilde{\epsilon}}{(1-\tilde{\epsilon})^2} \leq 1 + 4\tilde{\epsilon} = 1 + \epsilon_r. \tag{A.3}$$

where the inequality holds because $\epsilon_r \leq 0.5$ and $\tilde{\epsilon} = \frac{\epsilon_r}{4} \leq \frac{1}{8}$. Since we want (A.3) to hold with probability greater than $1 - e^{-t}$ it remains to require $\left|\mathcal{L}_r^{\text{corank}}\right|(1 + \frac{8}{\epsilon_r})^{d-r} e^{-\frac{C_M m\epsilon_r^2}{32}} \leq e^{-t}$. Using the fact that $(1 + \frac{8}{\epsilon_r}) \geq \frac{9}{\epsilon_r}$ and some arithmetics we get (20) and this completes the proof of the theorem.

We turn now to the proof of Theorem 3.8. Its proof is almost identical to the previous proof but with the difference that instead of r, $\mathcal{L}_r^{\text{corank}}$ and δ_r^{corank} we look at ℓ, $\mathcal{L}_\ell$ and δ_ℓ. In this case we do not know what is the dimension of the subspace that each cosupport implies. However, we can have a lower bound on it using $p-\ell$. Therefore, we use $B^{p-\ell}$ instead of B^{d-r}. This change provides us with a condition similar to (20) but with $p-\ell$ in the second coefficient instead of $d-r$. By using some arithmetics, noticing that the size of $\mathcal{L}_\ell$ is $\binom{p}{\ell}$ and using Stirling's formula for upper bounding it we get (21) and this completes the proof.

Appendix B. Proof of Lemma 6.6

Lemma 6.6: Consider the problem $\mathcal{P}$ and apply either AIHT or AHTP with a constant step size μ satisfying $\frac{1}{\mu} \geq 1+\delta_{2\ell-p}$ or an optimal step size. Then, at the t-th iteration, the following holds:

$$\begin{aligned}\left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2 - \left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2 \leq C_\ell\left(\|\mathbf{y}-\mathbf{M}\mathbf{x}\|_2^2 - \left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2\right) \\ + C_\ell\left(\frac{1}{\mu(1-\delta_{2\ell-p})}-1\right)\left\|\mathbf{M}(\mathbf{x}-\hat{\mathbf{x}}^{t-1})\right\|_2^2 + (C_\ell-1)\mu\sigma_{\mathbf{M}}^2\left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2.\end{aligned} \tag{B.1}$$

For the optimal step size the bound is achieved with the value $\mu = \frac{1}{1+\delta_{2\ell-p}}$.

Proof: We consider the AIHT algorithm first. We take similar steps to those taken in the proof of Lemma 3 in [29]. Since $\frac{1}{\mu} \geq 1+\delta_{2\ell-p}$, we have, from the $\mathbf{\Omega}$-RIP property of $\mathbf{M}$,

$$\left\|\mathbf{M}(\hat{\mathbf{x}}^t-\hat{\mathbf{x}}^{t-1})\right\|_2^2 \leq \frac{1}{\mu}\left\|\hat{\mathbf{x}}^t-\hat{\mathbf{x}}^{t-1}\right\|_2^2.$$

Thus,

$$\begin{aligned}\left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2 - \left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2 &= -2\langle\mathbf{M}(\hat{\mathbf{x}}^t-\hat{\mathbf{x}}^{t-1}), \mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1}\rangle + \left\|\mathbf{M}(\hat{\mathbf{x}}^t-\hat{\mathbf{x}}^{t-1})\right\|_2^2 \\ &\leq -2\langle\mathbf{M}(\hat{\mathbf{x}}^t-\hat{\mathbf{x}}^{t-1}), \mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1}\rangle + \frac{1}{\mu}\left\|\hat{\mathbf{x}}^t-\hat{\mathbf{x}}^{t-1}\right\|_2^2 \\ &= -2\langle\hat{\mathbf{x}}^t-\hat{\mathbf{x}}^{t-1}, \mathbf{M}^*(\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1})\rangle + \frac{1}{\mu}\left\|\hat{\mathbf{x}}^t-\hat{\mathbf{x}}^{t-1}\right\|_2^2 \\ &= -\mu\left\|\mathbf{M}^*(\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1})\right\|_2^2 + \frac{1}{\mu}\left\|\hat{\mathbf{x}}^t-\hat{\mathbf{x}}^{t-1}-\mu\mathbf{M}^*(\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1})\right\|_2^2.\end{aligned}$$

Note that by definition, $\hat{\mathbf{x}}^t = \mathbf{Q}_{\hat{\mathcal{S}}_\ell}\left(\hat{\mathbf{x}}^{t-1}+\mu\mathbf{M}^*(\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1})\right)$. Hence, by the C_ℓ-near optimality of the projection, we get

$$\left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2 - \left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2 \leq -\mu\left\|\mathbf{M}^*(\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1})\right\|_2^2 + \frac{C_\ell}{\mu}\left\|\mathbf{x}-\hat{\mathbf{x}}^{t-1}-\mu\mathbf{M}^*(\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1})\right\|_2^2. \tag{B.2}$$

Now note that

$$\begin{aligned}&\left\|\mathbf{x}-\hat{\mathbf{x}}^{t-1}-\mu\mathbf{M}^*(\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1})\right\|_2^2 \\ &\quad= \left\|\mathbf{x}-\hat{\mathbf{x}}^{t-1}\right\|_2^2 - 2\mu\langle\mathbf{M}(\mathbf{x}-\hat{\mathbf{x}}^{t-1}), \mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1}\rangle + \mu^2\left\|\mathbf{M}^*(\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1})\right\|_2^2 \\ &\quad\leq \frac{1}{1-\delta_{2\ell-p}}\left\|\mathbf{M}(\mathbf{x}-\hat{\mathbf{x}}^{t-1})\right\|_2^2 - 2\mu\langle\mathbf{M}(\mathbf{x}-\hat{\mathbf{x}}^{t-1}), \mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1}\rangle + \mu^2\left\|\mathbf{M}^*(\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1})\right\|_2^2 \\ &\quad= \frac{1}{1-\delta_{2\ell-p}}\left\|\mathbf{M}(\mathbf{x}-\hat{\mathbf{x}}^{t-1})\right\|_2^2 + \mu\left(\|\mathbf{y}-\mathbf{M}\mathbf{x}\|_2^2 - \left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2 - \left\|\mathbf{M}(\mathbf{x}-\hat{\mathbf{x}}^{t-1})\right\|_2^2\right) \\ &\qquad+ \mu^2\left\|\mathbf{M}^*(\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1})\right\|_2^2.\end{aligned}$$

Putting this into (B.2), we obtain the desired result for the AIHT algorithm.

We can check that the same holds true for the AHTP algorithm as follows: suppose that $\hat{\mathbf{x}}^{t-1}_{\text{AHTP}}$ is the $(t-1)$-st estimate from the AHTP algorithm. If we now initialize the AIHT algorithm with this estimate and obtain the next estimate $\hat{\mathbf{x}}^{t}_{\text{A}\tilde{\text{I}}\text{HT}}$, then the inequality of the lemma holds true with $\hat{\mathbf{x}}^{t}_{\text{A}\tilde{\text{I}}\text{HT}}$ and $\hat{\mathbf{x}}^{t-1}_{\text{AHTP}}$ in place of $\hat{\mathbf{x}}^{t}$ and $\hat{\mathbf{x}}^{t-1}$ respectively. On the other hand, from the algorithm description, we know that the t-th estimate $\hat{\mathbf{x}}^{t}_{\text{AHTP}}$ of the AHTP satisfies

$$\left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t}_{\text{AHTP}}\right\|_2^2 \leq \left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t}_{\text{A}\tilde{\text{I}}\text{HT}}\right\|_2^2.$$

This means that the result holds for the AHTP algorithm as well.

Using a similar argument for the optimal changing step size we note that it selects the cosupport that minimizes $\left\|\mathbf{M}\mathbf{x}-\mathbf{M}\hat{\mathbf{x}}^{t}\right\|_2^2$. Thus, for AIHT and AHTP we have that $\left\|\mathbf{M}\mathbf{x}-\mathbf{M}\hat{\mathbf{x}}^{t}_{\text{Opt}}\right\|_2^2 \leq \left\|\mathbf{M}\mathbf{x}-\mathbf{M}\hat{\mathbf{x}}^{t}_{\mu}\right\|_2^2$ for any value of μ, where $\hat{\mathbf{x}}^{t}_{\text{Opt}}$ and $\hat{\mathbf{x}}^{t}_{\mu}$ are the recovery results of AIHT or AHTP with an optimal changing step-size and a constant step-size μ respectively. This yields that any theoretical result for a constant step-size selection with a constant μ holds true also to the optimal changing-step size selection. In particular this is true also for $\mu = \frac{1}{1+\delta_{2\ell-p}}$. This choice is justified in the proof of Lemma 6.7. □

Appendix C. Proof of Lemma 6.7

Lemma 6.7: Suppose that the same conditions of Theorem 6.5 hold true. If $\left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2 \leq \eta^2\left\|\mathbf{e}\right\|_2^2$, then $\left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t}\right\|_2^2 \leq \eta^2\left\|\mathbf{e}\right\|_2^2$. Furthermore, if $\left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2 > \eta^2\left\|\mathbf{e}\right\|_2^2$, then

$$\left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t}\right\|_2^2 \leq c_4\left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2$$

where

$$c_4 := \left(1+\frac{1}{\eta}\right)^2\left(\frac{1}{\mu(1-\delta_{2\ell-p})}-1\right)C_\ell + (C_\ell-1)(\mu\sigma_{\mathbf{M}}^2-1) + \frac{C_\ell}{\eta^2} < 1.$$

Proof: First, suppose that $\left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2 > \eta^2\left\|\mathbf{e}\right\|_2^2$. From Lemma 6.6, we have

$$\begin{aligned}\left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t}\right\|_2^2 \quad &\leq \quad C_\ell\left\|\mathbf{y}-\mathbf{M}\mathbf{x}\right\|_2^2 + (C_\ell-1)(\mu\sigma_{\mathbf{M}}^2-1)\left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2 \\ &\quad + C_\ell\left(\frac{1}{\mu(1-\delta_{2\ell-p})}-1\right)\left\|\mathbf{M}(\mathbf{x}-\hat{\mathbf{x}}^{t-1})\right\|_2^2.\end{aligned} \tag{C.1}$$

Remark that all the coefficients in the above are positive because $1+\delta_{2\ell-p} \leq \frac{1}{\mu} \leq \sigma_{\mathbf{M}}^2$ and $C_\ell \geq 1$. Since $\mathbf{y}-\mathbf{M}\mathbf{x} = \mathbf{e}$, we note

$$\left\|\mathbf{y}-\mathbf{M}\mathbf{x}\right\|_2^2 < \frac{1}{\eta^2}\left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2$$

and, by the triangle inequality,

$$\left\|\mathbf{M}(\mathbf{x}-\hat{\mathbf{x}}^{t-1})\right\|_2 \leq \left\|\mathbf{y}-\mathbf{M}\mathbf{x}\right\|_2 + \left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2 < \left(1+\frac{1}{\eta}\right)\left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2.$$

Therefore, from (C.1),

$$\left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t}\right\|_2^2 < c_4\left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2.$$

This is the second part of the lemma.

Now, suppose that $\left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2 \leq \eta^2\left\|\mathbf{e}\right\|_2^2$. This time we have

$$\left\|\mathbf{M}(\mathbf{x}-\hat{\mathbf{x}}^{t-1})\right\|_2 \leq \left\|\mathbf{y}-\mathbf{M}\mathbf{x}\right\|_2 + \left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2 \leq (1+\eta)\left\|\mathbf{e}\right\|_2.$$

Applying this to (C.1), we obtain

$$\left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2 \le C_\ell \left\|\mathbf{e}\right\|_2^2 + (C_\ell-1)(\mu\sigma_{\mathbf{M}}^2-1)\eta^2\left\|\mathbf{e}\right\|_2^2 + C_\ell\left(\frac{1}{\mu(1-\delta_{2\ell-p})}-1\right)(1+\eta)^2\left\|\mathbf{e}\right\|_2^2$$
$$= \left(C_\ell + (C_\ell-1)(\mu\sigma_M^2-1)\eta^2 + C_\ell\left(\frac{1}{\mu(1-\delta_{2\ell-p})}-1\right)(1+\eta)^2\right)\left\|\mathbf{e}\right\|_2^2 = c_4\eta^2\left\|\mathbf{e}\right\|_2^2.$$

Thus, the proof is complete as soon as we show $c_4 < 1$, or $c_4 - 1 < 0$.

To see $c_4 - 1 < 0$, we first note that it is equivalent to–all the subscripts are dropped from here on for simplicity of notation–

$$\frac{1}{\mu^2} - \frac{2(1-\delta)}{1+\frac{1}{\eta}}\frac{1}{\mu} + \frac{(C-1)\sigma^2(1-\delta)}{C\left(1+\frac{1}{\eta}\right)^2} < 0,$$

or

$$\frac{1}{\mu^2} - 2(1-\delta)b_1\frac{1}{\mu} + (1-\delta)^2 b_2 < 0.$$

Solving this quadratic equation in $\frac{1}{\mu}$, we want

$$(1-\delta)\left(b_1 - \sqrt{b_1^2-b_2}\right) < \frac{1}{\mu} < (1-\delta)\left(b_1 + \sqrt{b_1^2-b_2}\right).$$

Such μ exists only when $\frac{b_2}{b_1^2} < 1$. Furthermore, we have already assumed $1+\delta \le \frac{1}{\mu}$ and we know $(1-\delta)\left(b_1 - \sqrt{b_1^2-b_2}\right) < 1+\delta$, and hence the condition we require is

$$1+\delta \le \frac{1}{\mu} < (1-\delta)\left(b_1 + \sqrt{b_1^2-b_2}\right),$$

which is what we desired to prove.

As we have seen in Lemma 6.6, for changing optimal step-size selection, (49) holds for any value of μ that satisfies the above conditions. Thus, in the bound of changing optimal step-size we put a value of μ that minimizes c_4. This minimization result with $\frac{1}{\mu} = \sqrt{b_2}(1-\delta_{2\ell-p})$. However, since we need $\frac{1}{\mu} \ge 1+\delta_{2\ell-p}$ and have that $\sqrt{b_2}(1-\delta_{2\ell-p}) < b_1(1-\delta_{2\ell-p}) < 1+\delta_{2\ell-p}$ we set $\frac{1}{\mu} = 1+\delta_{2\ell-p}$ in c_4 for the bound in optimal changing step-size case. □

Appendix D. Proof of Lemma 6.10

Lemma 6.10: Consider the problem $\mathcal{P}$ and apply ACoSaMP with $a = \frac{2\ell-p}{\ell}$. For each iteration we have

$$\left\|\mathbf{x}-\mathbf{w}\right\|_2 \quad \le \quad \frac{1}{\sqrt{1-\delta_{4\ell-3p}^2}}\left\|\mathbf{P}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2 + \frac{\sqrt{1+\delta_{3\ell-2p}}}{1-\delta_{4\ell-3p}}\left\|\mathbf{e}\right\|_2.$$

Proof: Since $\mathbf{w}$ is the minimizer of $\left\|\mathbf{y}-\mathbf{M}\mathbf{v}\right\|_2^2$ with the constraint $\mathbf{\Omega}_{\tilde{\Lambda}^t}\mathbf{v} = 0$, then

$$\langle \mathbf{M}\mathbf{w}-\mathbf{y}, \mathbf{M}\mathbf{u}\rangle = 0, \tag{D.1}$$

for any vector $\mathbf{u}$ such that $\mathbf{\Omega}_{\tilde{\Lambda}^t}\mathbf{u} = 0$. Substituting $\mathbf{y} = \mathbf{M}\mathbf{x}+\mathbf{e}$ and moving terms from the LHS to the RHS gives

$$\langle \mathbf{w}-\mathbf{x}, \mathbf{M}^*\mathbf{M}\mathbf{u}\rangle = \langle \mathbf{e}, \mathbf{M}\mathbf{u}\rangle, \tag{D.2}$$

where $\mathbf{u}$ is a vector satisfying $\mathbf{\Omega}_{\tilde{\Lambda}^t}\mathbf{u} = 0$. Turning to look at $\left\|\mathbf{Q}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2^2$ and using (D.2) with $\mathbf{u} = \mathbf{Q}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})$, we have

$$\begin{aligned}
&\left\|\mathbf{Q}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2^2 = \langle \mathbf{x}-\mathbf{w}, \mathbf{Q}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\rangle \\
&= \langle \mathbf{x}-\mathbf{w}, (\mathbf{I}-\mathbf{M}^*\mathbf{M})\mathbf{Q}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\rangle - \langle \mathbf{e}, \mathbf{M}\mathbf{Q}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\rangle \\
&\le \|\mathbf{x}-\mathbf{w}\|_2 \left\|\mathbf{Q}_{\Lambda\cap\tilde{\Lambda}^t}(\mathbf{I}-\mathbf{M}^*\mathbf{M})\mathbf{Q}_{\tilde{\Lambda}^t}\right\|_2 \left\|\mathbf{Q}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2 + \|\mathbf{e}\|_2 \left\|\mathbf{M}\mathbf{Q}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2 \\
&\le \delta_{4\ell-3p}\|\mathbf{x}-\mathbf{w}\|_2 \left\|\mathbf{Q}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2 + \|\mathbf{e}\|_2 \sqrt{1+\delta_{3\ell-2p}} \left\|\mathbf{Q}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2 .
\end{aligned} \tag{D.3}$$

where the first inequality follows from the Cauchy-Schwartz inequality, the projection property that $\mathbf{Q}_{\tilde{\Lambda}^t} = \mathbf{Q}_{\tilde{\Lambda}^t}\mathbf{Q}_{\tilde{\Lambda}^t}$ and the fact that $\mathbf{x}-\mathbf{w} = \mathbf{Q}_{\Lambda\cap\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})$. The last inequality is due to the $\mathbf{\Omega}$-RIP properties, Corollary 3.6 and that according to Table 1 $|\tilde{\Lambda}^t| \ge 3\ell - 2p$ and $|\Lambda\cap\tilde{\Lambda}^t| \ge 4\ell - 3p$. After simplification of (D.3) by $\left\|\mathbf{Q}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2$ we have

$$\left\|\mathbf{Q}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2 \le \delta_{4\ell-3p}\|\mathbf{x}-\mathbf{w}\|_2 + \sqrt{1+\delta_{3\ell-2p}}\,\|\mathbf{e}\|_2 .$$

Utilizing the last inequality with the fact that $\|\mathbf{x}-\mathbf{w}\|_2^2 = \left\|\mathbf{P}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2^2 + \left\|\mathbf{Q}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2^2$ gives

$$\|\mathbf{x}-\mathbf{w}\|_2^2 \le \left\|\mathbf{P}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2^2 + \left(\delta_{4\ell-3p}\|\mathbf{x}-\mathbf{w}\|_2 + \sqrt{1+\delta_{3\ell-2p}}\,\|\mathbf{e}\|_2\right)^2 . \tag{D.4}$$

By moving all terms to the LHS we get a quadratic function of $\|\mathbf{x}-\mathbf{w}\|_2$. Thus, $\|\mathbf{x}-\mathbf{w}\|_2$ is bounded from above by the larger root of that function; this with a few simple algebraic steps gives the inequality in (58). □

Appendix E. Proof of Lemma 6.11

Lemma 6.11: Consider the problem $\mathcal{P}$ and apply ACoSaMP with $a = \frac{2\ell-p}{\ell}$. For each iteration we have

$$\left\|\mathbf{x}-\hat{\mathbf{x}}^t\right\|_2 \le \rho_1 \left\|\mathbf{P}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2 + \eta_1\|\mathbf{e}\|_2 ,$$

where η_1 and ρ_1 are the same constants as in Theorem 6.8.

Proof: We start with the following observation

$$\left\|\mathbf{x}-\hat{\mathbf{x}}^t\right\|_2^2 = \left\|\mathbf{x}-\mathbf{w}+\mathbf{w}-\hat{\mathbf{x}}^t\right\|_2^2 = \|\mathbf{x}-\mathbf{w}\|_2^2 + \left\|\hat{\mathbf{x}}^t-\mathbf{w}\right\|_2^2 + 2(\mathbf{x}-\mathbf{w})^*(\mathbf{w}-\hat{\mathbf{x}}^t), \tag{E.1}$$

and turn to bound the second and last terms in the RHS. For the second term, using the fact that $\hat{\mathbf{x}}^t = \mathbf{Q}_{\hat{\mathcal{S}}_\ell(\mathbf{w})}\mathbf{w}$ with (24) gives

$$\left\|\hat{\mathbf{x}}^t-\mathbf{w}\right\|_2^2 \le C_\ell\|\mathbf{x}-\mathbf{w}\|_2^2 . \tag{E.2}$$

For bounding the last term, we look at its absolute value and use (D.2) with $\mathbf{u} = \mathbf{w}-\hat{\mathbf{x}}^t = \mathbf{Q}_{\tilde{\Lambda}^t}(\mathbf{w}-\hat{\mathbf{x}}^t)$. This leads to

$$\left|(\mathbf{x}-\mathbf{w})^*(\mathbf{w}-\hat{\mathbf{x}}^t)\right| = \left|(\mathbf{x}-\mathbf{w})^*(\mathbf{I}-\mathbf{M}^*\mathbf{M})(\mathbf{w}-\hat{\mathbf{x}}^t) - \mathbf{e}^*\mathbf{M}(\mathbf{w}-\hat{\mathbf{x}}^t)\right| .$$

By using the triangle and Cauchy-Schwartz inequalities with the fact that $\mathbf{x}-\mathbf{w} = \mathbf{Q}_{\Lambda\cap\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})$ and $\mathbf{w}-\hat{\mathbf{x}}^t = \mathbf{Q}_{\tilde{\Lambda}^t}(\mathbf{w}-\hat{\mathbf{x}}^t)$ we have

$$\begin{aligned}
\left|(\mathbf{x}-\mathbf{w})^*(\mathbf{w}-\hat{\mathbf{x}}^t)\right| &\le \|\mathbf{x}-\mathbf{w}\|_2 \left\|\mathbf{Q}_{\Lambda\cap\tilde{\Lambda}^t}(\mathbf{I}-\mathbf{M}^*\mathbf{M})\mathbf{Q}_{\tilde{\Lambda}^t}\right\|_2 \left\|\mathbf{w}-\hat{\mathbf{x}}^t\right\|_2 + \|\mathbf{e}\|_2 \left\|\mathbf{M}(\mathbf{w}-\hat{\mathbf{x}}^t)\right\|_2 \\
&\le \delta_{4\ell-3p}\|\mathbf{x}-\mathbf{w}\|_2 \left\|\mathbf{w}-\hat{\mathbf{x}}^t\right\|_2 + \sqrt{1+\delta_{3\ell-2p}}\,\|\mathbf{e}\|_2 \left\|\mathbf{w}-\hat{\mathbf{x}}^t\right\|_2 ,
\end{aligned} \tag{E.3}$$

where the last inequality is due to the $\mathbf{\Omega}$-RIP definition and Corollary 3.6.

By substituting (E.2) and (E.3) into (E.1) we have

$$\begin{aligned}
\left\|\mathbf{x}-\hat{\mathbf{x}}^t\right\|_2^2 &\leq (1+C_\ell)\left\|\mathbf{x}-\mathbf{w}\right\|_2^2 + 2\delta_{4\ell-3p}\sqrt{C_\ell}\left\|\mathbf{x}-\mathbf{w}\right\|_2^2 + 2\sqrt{1+\delta_{3\ell-2p}}\sqrt{C_\ell}\left\|\mathbf{e}\right\|_2\left\|\mathbf{x}-\mathbf{w}\right\|_2 \qquad \text{(E.4)}\\
&\leq \left((1+2\delta_{4\ell-3p}\sqrt{C_\ell}+C_\ell)\left\|\mathbf{x}-\mathbf{w}\right\|_2 + 2\sqrt{(1+\delta_{3\ell-2p})C_\ell}\left\|\mathbf{e}\right\|_2\right)\left\|\mathbf{x}-\mathbf{w}\right\|_2\\
&\leq \frac{1+2\delta_{4\ell-3p}\sqrt{C_\ell}+C_\ell}{1-\delta_{4\ell-3p}^2}\left\|\mathbf{P}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2^2\\
&+\frac{2\sqrt{1+\delta_{3\ell-2p}}(1+(1+\delta_{4\ell-3p})\sqrt{C_\ell}+C_\ell)}{(1-\delta_{4\ell-3p})\sqrt{1-\delta_{4\ell-3p}^2}}\left\|\mathbf{P}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2\left\|\mathbf{e}\right\|_2 + \frac{(1+\delta_{3\ell-2p})(1+2\sqrt{C_\ell}+C_\ell)}{(1-\delta_{4\ell-3p})^2}\left\|\mathbf{e}\right\|_2^2\\
&\leq \left(\frac{\sqrt{1+2\delta_{4\ell-3p}\sqrt{C_\ell}+C_\ell}}{\sqrt{1-\delta_{4\ell-3p}^2}}\left\|\mathbf{P}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2 + \frac{\sqrt{\frac{2+C_\ell}{1+C_\ell}+2\sqrt{C_\ell}+C_\ell}\sqrt{1+\delta_{3\ell-2p}}}{1-\delta_{4\ell-3p}}\left\|\mathbf{e}\right\|_2\right)^2,
\end{aligned}$$

where for the second inequality we use the fact that $\delta_{4\ell-3p}\leq 1$ combined with the inequality of Lemma 6.10, and for the last inequality we use the fact that $(1+(1+\delta_{4\ell-3p})\sqrt{C_\ell}+C_\ell)^2 \leq (1+2\delta_{4\ell-3p}\sqrt{C_\ell}+C_\ell)(\frac{2+C_\ell}{1+C_\ell}+2\sqrt{C_\ell}+C_\ell)$ together with a few algebraic steps. Taking square-root on both sides of (E.4) provides the desired result. □

Appendix F. Proof of Lemma 6.12

Lemma 6.12: Consider the problem $\mathcal{P}$ and apply ACoSaMP with $a=\frac{2\ell-p}{\ell}$. if

$$C_{2\ell-p} < \frac{\sigma_{\mathbf{M}}^2(1+\gamma)^2}{\sigma_{\mathbf{M}}^2(1+\gamma)^2-1},$$

then there exists $\tilde{\delta}_{\text{ACoSaMP}}(C_{2\ell-p},\sigma_{\mathbf{M}}^2,\gamma)>0$ such that for any $\delta_{2\ell-p}<\tilde{\delta}_{\text{ACoSaMP}}(C_{2\ell-p},\sigma_{\mathbf{M}}^2,\gamma)$

$$\left\|\mathbf{P}_{\tilde{\Lambda}^t}(\mathbf{x}-\mathbf{w})\right\|_2 \leq \eta_2\left\|\mathbf{e}\right\|_2 + \rho_2\left\|\mathbf{x}-\hat{\mathbf{x}}^{t-1}\right\|_2.$$

The constants η_2 and ρ_2 are as defined in Theorem 6.8.

In the proof of the lemma we use the following Proposition.

Proposition E.1: For any two given vectors $\mathbf{x}_1,\mathbf{x}_2$ and any constant $c>0$ it holds that

$$\left\|\mathbf{x}_1+\mathbf{x}_2\right\|_2^2 \leq (1+c)\left\|\mathbf{x}_1\right\|_2^2 + \left(1+\frac{1}{c}\right)\left\|\mathbf{x}_2\right\|^2. \qquad \text{(F.1)}$$

The proof of the proposition is immediate using the inequality of arithmetic and geometric means. We turn to the proof of the lemma.

Proof: Looking at the step of finding new cosupport elements one can observe that $\mathbf{Q}_{\Lambda_\Delta}$ is a near optimal projection for $\mathbf{M}^*\mathbf{y}_{\text{resid}}^{t-1}=\mathbf{M}^*(\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1})$ with a constant $C_{2\ell-p}$. The fact that $\left|\hat{\Lambda}^{t-1}\cap\Lambda\right|\geq 2\ell-p$ combined with (24) gives

$$\left\|(\mathbf{I}-\mathbf{Q}_{\Lambda_\Delta})\mathbf{M}^*(\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1})\right\|_2^2 \leq C_{2\ell-p}\left\|(\mathbf{I}-\mathbf{Q}_{\hat{\Lambda}^{t-1}\cap\Lambda})\mathbf{M}^*(\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1})\right\|_2^2.$$

Using simple projection properties and the fact that $\tilde{\Lambda}^t\subseteq\Lambda_\Delta$ with $\mathbf{z}=\mathbf{M}^*(\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^{t-1})$ we have

$$\begin{aligned}
\left\|\mathbf{Q}_{\tilde{\Lambda}^t}\mathbf{z}\right\|_2^2 \geq \left\|\mathbf{Q}_{\Lambda_\Delta}\mathbf{z}\right\|_2^2 &= \left\|\mathbf{z}\right\|_2^2 - \left\|(\mathbf{I}-\mathbf{Q}_{\Lambda_\Delta})\mathbf{z}\right\|_2^2 \geq \left\|\mathbf{z}\right\|_2^2 - C_{2\ell-p}\left\|(\mathbf{I}-\mathbf{Q}_{\hat{\Lambda}^{t-1}\cap\Lambda})\mathbf{z}\right\|_2^2 \qquad \text{(F.2)}\\
&= \left\|\mathbf{z}\right\|_2^2 - C_{2\ell-p}\left(\left\|\mathbf{z}\right\|_2^2 - \left\|\mathbf{Q}_{\hat{\Lambda}^{t-1}\cap\Lambda}\mathbf{z}\right\|_2^2\right) = C_{2\ell-p}\left\|\mathbf{Q}_{\hat{\Lambda}^{t-1}\cap\Lambda}\mathbf{z}\right\|_2^2 - (C_{2\ell-p}-1)\left\|\mathbf{z}\right\|_2^2.
\end{aligned}$$

We turn to bound the LHS of (F.2) from above. Noticing that $\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{e}$ and using (F.1) with a constant $\gamma_1 > 0$ gives

$$\left\|\mathbf{Q}_{\tilde{\Lambda}^t}\mathbf{M}^*(\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1})\right\|_2^2 \leq \left(1 + \frac{1}{\gamma_1}\right)\left\|\mathbf{Q}_{\tilde{\Lambda}^t}\mathbf{M}^*\mathbf{e}\right\|_2^2 + (1+\gamma_1)\left\|\mathbf{Q}_{\tilde{\Lambda}^t}\mathbf{M}^*\mathbf{M}(\mathbf{x} - \hat{\mathbf{x}}^{t-1})\right\|_2^2. \tag{F.3}$$

Using (F.1) again, now with a constant $\alpha > 0$, we have

$$\begin{aligned}\left\|\mathbf{Q}_{\tilde{\Lambda}^t}\mathbf{M}^*\mathbf{M}(\mathbf{x} - \hat{\mathbf{x}}^{t-1})\right\|_2^2 &\leq (1+\alpha)\left\|\mathbf{Q}_{\tilde{\Lambda}^t}(\mathbf{x} - \hat{\mathbf{x}}^{t-1})\right\|_2^2 + \left(1 + \frac{1}{\alpha}\right)\left\|\mathbf{Q}_{\tilde{\Lambda}^t}(\mathbf{I} - \mathbf{M}^*\mathbf{M})(\mathbf{x} - \hat{\mathbf{x}}^{t-1})\right\|_2^2 \\ &\leq (1+\alpha)\left\|\mathbf{x} - \hat{\mathbf{x}}^{t-1}\right\|_2^2 - (1+\alpha)\left\|\mathbf{P}_{\tilde{\Lambda}^t}(\mathbf{x} - \hat{\mathbf{x}}^{t-1})\right\|_2^2 + \left(1 + \frac{1}{\alpha}\right)\left\|\mathbf{Q}_{\tilde{\Lambda}^t}(\mathbf{I} - \mathbf{M}^*\mathbf{M})(\mathbf{x} - \hat{\mathbf{x}}^{t-1})\right\|_2^2.\end{aligned} \tag{F.4}$$

Putting (F.4) into (F.3) and using (18) and Corollary 3.3 gives

$$\begin{aligned}\left\|\mathbf{Q}_{\tilde{\Lambda}^t}\mathbf{M}^*(\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1})\right\|_2^2 \leq{}& \frac{(1+\gamma_1)(1+\delta_{3\ell-2p})}{\gamma_1}\|\mathbf{e}\|_2^2 - (1+\alpha)(1+\gamma_1)\left\|\mathbf{P}_{\tilde{\Lambda}^t}(\mathbf{x} - \hat{\mathbf{x}}^{t-1})\right\|_2^2 \\ &+ \left(1 + \alpha + \delta_{4\ell-3p} + \frac{\delta_{4\ell-3p}}{\alpha}\right)(1+\gamma_1)\left\|\mathbf{x} - \hat{\mathbf{x}}^{t-1}\right\|_2^2.\end{aligned} \tag{F.5}$$

We continue with bounding the RHS of (F.2) from below. For the first element of the RHS we use an altered version of (F.1) with a constant $\gamma_2 > 0$ and have

$$\left\|\mathbf{Q}_{\hat{\Lambda}^{t-1}\cap\Lambda}\mathbf{M}^*(\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1})\right\|_2^2 \geq \frac{1}{1+\gamma_2}\left\|\mathbf{Q}_{\hat{\Lambda}^{t-1}\cap\Lambda}\mathbf{M}^*\mathbf{M}(\mathbf{x} - \hat{\mathbf{x}}^{t-1})\right\|_2^2 - \frac{1}{\gamma_2}\left\|\mathbf{Q}_{\hat{\Lambda}^{t-1}\cap\Lambda}\mathbf{M}^*\mathbf{e}\right\|_2^2. \tag{F.6}$$

Using the altered form again, for the first element in the RHS of (F.6), with a constant $\beta > 0$ gives

$$\left\|\mathbf{Q}_{\hat{\Lambda}^{t-1}\cap\Lambda}\mathbf{M}^*\mathbf{M}(\mathbf{x} - \hat{\mathbf{x}}^{t-1})\right\|_2^2 \geq \frac{1}{1+\beta}\left\|\mathbf{x} - \hat{\mathbf{x}}^{t-1}\right\|_2^2 - \frac{1}{\beta}\left\|\mathbf{Q}_{\hat{\Lambda}^{t-1}\cap\Lambda}(\mathbf{M}^*\mathbf{M} - \mathbf{I})(\mathbf{x} - \hat{\mathbf{x}}^{t-1})\right\|_2^2. \tag{F.7}$$

Putting (F.7) in (F.6) and using the RIP properties and (18) provide

$$\left\|\mathbf{Q}_{\hat{\Lambda}^{t-1}\cap\Lambda}\mathbf{M}^*(\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1})\right\|_2^2 \geq \left(\frac{1}{1+\beta} - \frac{\delta_{2\ell-p}}{\beta}\right)\frac{1}{1+\gamma_2}\left\|\mathbf{x} - \hat{\mathbf{x}}^{t-1}\right\|_2^2 - \frac{(1+\delta_{2\ell-p})}{\gamma_2}\|\mathbf{e}\|_2^2. \tag{F.8}$$

Using (F.1), with a constant $\gamma_3 > 0$, (9), and some basic algebraic steps we have for the second element in the RHS of (F.2)

$$\begin{aligned}\left\|\mathbf{M}^*(\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1})\right\|_2^2 &\leq (1+\gamma_3)\left\|\mathbf{M}^*\mathbf{M}(\mathbf{x} - \hat{\mathbf{x}}^{t-1})\right\|_2^2 + \left(1 + \frac{1}{\gamma_3}\right)\|\mathbf{M}^*\mathbf{e}\|_2^2 \\ &\leq (1+\gamma_3)(1+\delta_{2\ell-p})\sigma_{\mathbf{M}}^2\left\|(\mathbf{x} - \hat{\mathbf{x}}^{t-1})\right\|_2^2 + \left(1 + \frac{1}{\gamma_3}\right)\sigma_{\mathbf{M}}^2\|\mathbf{e}\|_2^2.\end{aligned} \tag{F.9}$$

By combining (F.5), (F.8) and (F.9) with (F.2) we have

$$\begin{aligned}(1+\alpha)(1+\gamma_1)\left\|\mathbf{P}_{\tilde{\Lambda}^t}(\mathbf{x} - \hat{\mathbf{x}}^{t-1})\right\|_2^2 \leq{}& \frac{(1+\gamma_1)(1+\delta_{3\ell-2p})}{\gamma_1}\|\mathbf{e}\|_2^2 + C_{2\ell-p}\frac{(1+\delta_{2\ell-p})}{\gamma_2}\|\mathbf{e}\|_2^2 \\ &+(C_{2\ell-p}-1)\left(1 + \frac{1}{\gamma_3}\right)\sigma_{\mathbf{M}}^2\|\mathbf{e}\|_2^2 + \left(1 + \alpha + \delta_{4\ell-3p} + \frac{\delta_{4\ell-3p}}{\alpha}\right)(1+\gamma_1)\left\|\mathbf{x} - \hat{\mathbf{x}}^{t-1}\right\|_2^2 \\ &+(C_{2\ell-p}-1)(1+\gamma_3)(1+\delta_{2\ell-p})\sigma_{\mathbf{M}}^2\left\|(\mathbf{x} - \hat{\mathbf{x}}^{t-1})\right\|_2^2 - C_{2\ell-p}\left(\frac{1}{1+\beta} - \frac{\delta_{2\ell-p}}{\beta}\right)\frac{1}{1+\gamma_2}\left\|\mathbf{x} - \hat{\mathbf{x}}^{t-1}\right\|_2^2.\end{aligned} \tag{F.10}$$

Dividing both sides by $(1+\alpha)(1+\gamma_1)$ and gathering coefficients give

$$\begin{aligned}\left\|\mathbf{P}_{\tilde{\Lambda}^t}(\mathbf{x}-\hat{\mathbf{x}}^{t-1})\right\|_2^2 \leq &\left(\frac{1+\delta_{3\ell-2p}}{\gamma_1(1+\alpha)}+\frac{(1+\delta_{2\ell-p})C_{2\ell-p}}{\gamma_2(1+\alpha)(1+\gamma_1)}+\frac{(C_{2\ell-p}-1)(1+\gamma_3)\sigma_{\mathbf{M}}^2}{(1+\alpha)(1+\gamma_1)\gamma_3}\right)\|\mathbf{e}\|_2^2 \\ &+\left(1+\frac{\delta_{4\ell-3p}}{\alpha}+\frac{(C_{2\ell-p}-1)(1+\gamma_3)(1+\delta_{2\ell-p})\sigma_{\mathbf{M}}^2}{(1+\alpha)(1+\gamma_1)}\right. \\ &\left.-\frac{C_{2\ell-p}}{(1+\alpha)(1+\gamma_1)(1+\gamma_2)}\left(\frac{1}{1+\beta}-\frac{\delta_{2\ell-p}}{\beta}\right)\right)\left\|\mathbf{x}-\hat{\mathbf{x}}^{t-1}\right\|_2^2.\end{aligned} \tag{F.11}$$

The smaller the coefficient of $\left\|\mathbf{x}-\hat{\mathbf{x}}^{t-1}\right\|_2^2$, the better convergence guarantee we obtain. Thus, we choose $\beta=\frac{\sqrt{\delta_{2\ell-p}}}{1-\sqrt{\delta_{2\ell-p}}}$ and $\alpha=\frac{\sqrt{\delta_{4\ell-3p}}}{\sqrt{\frac{C_{2\ell-p}}{(1+\gamma_1)(1+\gamma_2)}\left(1-\sqrt{\delta_{2\ell-p}}\right)^2-\frac{(C_{2\ell-p}-1)(1+\gamma_3)(1+\delta_{2\ell-p})\sigma_{\mathbf{M}}^2}{1+\gamma_1}}-\sqrt{\delta_{4\ell-3p}}}$ so that the coefficient is minimized. The values of $\gamma_1,\gamma_2,\gamma_3$ provide a tradeoff between the convergence rate and the size of the noise coefficient. For smaller values we get better convergence rate but higher amplification of the noise. We make no optimization on their values and choose them to be $\gamma_1=\gamma_2=\gamma_3=\gamma$ for an appropriate $\gamma>0$. Thus, the above yields

$$\begin{aligned}\left\|\mathbf{P}_{\tilde{\Lambda}^t}(\mathbf{x}-\hat{\mathbf{x}}^{t-1})\right\|_2^2 \leq &\left(\frac{1+\delta_{3\ell-2p}}{\gamma(1+\alpha)}+\frac{(1+\delta_{2\ell-p})C_{2\ell-p}}{\gamma(1+\alpha)(1+\gamma)}+\frac{(C_{2\ell-p}-1)(1+\gamma)\sigma_{\mathbf{M}}^2}{(1+\alpha)(1+\gamma)\gamma}\right)\|\mathbf{e}\|_2^2 \\ &+\left(1-\left(\sqrt{\delta_{4\ell-3p}}-\sqrt{\frac{C_{2\ell-p}}{(1+\gamma)^2}\left(1-\sqrt{\delta_{2\ell-p}}\right)^2-(C_{2\ell-p}-1)(1+\delta_{2\ell-p})\sigma_{\mathbf{M}}^2}\right)^2\right)\left\|\mathbf{x}-\hat{\mathbf{x}}^{t-1}\right\|_2^2.\end{aligned} \tag{F.12}$$

Since $\mathbf{P}_{\tilde{\Lambda}^t}\mathbf{w}=\mathbf{P}_{\tilde{\Lambda}^t}\hat{\mathbf{x}}^{t-1}=0$ the above inequality holds also for $\left\|\mathbf{P}_{\tilde{\Lambda}^t}(\mathbf{x}-\hat{\mathbf{x}}^{t-1})\right\|_2^2$. Inequality (61) follows since the right-hand side of (F.12) is smaller than the square of the right-hand side of (61).

Before ending the proof, we notice that ρ_2, the coefficient of $\left\|\mathbf{x}-\hat{\mathbf{x}}^{t-1}\right\|_2^2$ is defined only when

$$(C_{2\ell-p}-1)(1+\delta_{2\ell-p})\sigma_{\mathbf{M}}^2 \leq \frac{C_{2\ell-p}}{(1+\gamma)^2}\left(1-\sqrt{\delta_{2\ell-p}}\right)^2. \tag{F.13}$$

First we notice that since $1+\delta_{2\ell-p}\geq\left(1-\sqrt{\delta_{2\ell-p}}\right)^2$ a necessary condition for (F.13) to hold is $(C_{2\ell-p}-1)\sigma_{\mathbf{M}}^2<\frac{C_{2\ell-p}}{(1+\gamma)^2}$ which is equivalent to (60). By moving the terms in the RHS to the LHS we get a quadratic function of $\sqrt{\delta_{2\ell-p}}$. The condition in (60) guarantees that its constant term is smaller than zero and thus there exists a positive $\delta_{2\ell-p}$ for which the function is smaller than zero. Therefore, for any $\delta_{2\ell-p}<\tilde{\delta}_{\text{ACoSaMP}}(C_{2\ell-p},\sigma_{\mathbf{M}}^2,\gamma)$ (F.13) holds, where $\tilde{\delta}_{\text{ACoSaMP}}(C_{2\ell-p},\sigma_{\mathbf{M}}^2,\gamma)>0$ is the square of the positive solution of the quadratic function.

□

Acknowledgment

The authors would like to thank Jalal Fadili for fruitful discussion, and the unknown reviewers for the important remarks that helped to improved the shape of the paper. Without both of them, the examples of the optimal projections would not have appeared in the paper. This research was supported by New York Metropolitan Research Fund. R. Giryes is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship. This work was supported in part by the EU FP7, SMALL project under FET-Open grant number 225913, and EPSRC grants EP/J015180/1 and EP/F039697/1. R. Gribonval acknowledges the support of the European Research Council, PLEASE project, under grant ERC-StG- 2011-277906. MED acknowledges support of his position from the Scottish Funding Council and their support of the Joint Research Institute with the Heriot-Watt University as a component part of the Edinburgh Research Partnership.

References

[1] Y. Lu, M. Do, A theory for sampling signals from a union of subspaces, IEEE Trans. Signal Process. 56 (6) (2008) 2334 –2345.

[2] D. Donoho, M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization, Proc. Nat. Aca. Sci. 100 (5) (2003) 2197–2202.

[3] R. Gribonval, M. Nielsen, Sparse representations in unions of bases, IEEE Trans. Inf. Theory. 49 (12) (2003) 3320–3325.

[4] G. Davis, S. Mallat, M. Avellaneda, Adaptive greedy approximations, Constructive Approximation 13 (1997) 57–98.

[5] E. J. Candès, T. Tao, Near-optimal signal recovery from random projections: Universal encoding strategies?, IEEE Trans. Inf. Theory. 52 (12) (2006) 5406 –5425.

[6] S. Foucart, Sparse recovery algorithms: sufficient conditions in terms of restricted isometry constants, in: Approximation Theory XIII, Springer Proceedings in Mathematics, 2010, pp. 65–77.

[7] Q. Mo, S. Li, New bounds on the restricted isometry constant, Applied and Computational Harmonic Analysis 31 (3) (2011) 460 – 468.

[8] H. Rauhut, K. Schnass, P. Vandergheynst, Compressed sensing and redundant dictionaries, IEEE Trans. Inf. Theory. 54 (5) (2008) 2210 –2219.

[9] Y. Pati, R. Rezaiifar, P. Krishnaprasad, Orthonormal matching pursuit : recursive function approximation with applications to wavelet decomposition, in: Proceedings of the 27^{th} Annual Asilomar Conf. on Signals, Systems and Computers, 1993.

[10] S. Mallat, Z. Zhang, Matching pursuits with time-frequency dictionaries, IEEE Trans. Signal Process. 41 (1993) 3397–3415.

[11] D. Needell, J. Tropp, CoSaMP: Iterative signal recovery from incomplete and inaccurate samples, Applied and Computational Harmonic Analysis 26 (3) (2009) 301 – 321.

[12] W. Dai, O. Milenkovic, Subspace pursuit for compressive sensing signal reconstruction, IEEE Trans. Inf. Theory. 55 (5) (2009) 2230 –2249.

[13] T. Blumensath, M. Davies, Iterative hard thresholding for compressed sensing, Applied and Computational Harmonic Analysis 27 (3) (2009) 265 – 274.

[14] S. Foucart, Hard thresholding pursuit: an algorithm for compressive sensing, SIAM J. Numer. Anal. 49 (6) (2011) 2543–2563.

[15] R. Giryes, M. Elad, RIP-based near-oracle performance guarantees for SP, CoSaMP, and IHT, IEEE Trans. Signal Process. 60 (3) (2012) 1465–1468.

[16] R. Garg, R. Khandekar, Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property, in: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, ACM, New York, NY, USA, 2009, pp. 337–344.

[17] T. Zhang, Sparse recovery with orthogonal matching pursuit under RIP, IEEE Trans. Inf. Theory. 57 (9) (2011) 6215 –6221.

[18] S. Nam, M. Davies, M. Elad, R. Gribonval, The cosparse analysis model and algorithms, Applied and Computational Harmonic Analysis.

[19] S. Nam, M. Davies, M. Elad, R. Gribonval, Cosparse analysis modeling - uniqueness and algorithms, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011.

[20] M. Elad, P. Milanfar, R. Rubinstein, Analysis versus synthesis in signal priors, Inverse Problems 23 (3) (2007) 947–968.

[21] E. J. Candès, Y. C. Eldar, D. Needell, P. Randall, Compressed sensing with coherent and redundant dictionaries, Applied and Computational Harmonic Analysis 31 (1) (2011) 59 – 73.

[22] S. Vaiter, G. Peyrè, C. Dossal, J. Fadili, Robust sparse analysis regularization, submitted to IEEE Trans. on Information Theory.

[23] S. Nam, M. Davies, M. Elad, R. Gribonval, Cosparse analysis modeling, in: 9th International Conference on SamplingTheory and Applications (sampta-2011), Singapore, 2011.

[24] I. Daubechies, R. DeVore, M. Fornasier, C. S. Güntürk, Iteratively reweighted least squares minimization for sparse recovery, Communications on Pure and Applied Mathematics 63 (1) (2010) 1–38.

[25] R. Rubinstein, T. Peleg, M. Elad, Analysis K-SVD: A dictionary learning algorithm for the analysis sparse model, submitted to IEEE Trans. on Signal Processing.

[26] T. Peleg, M. Elad, Performance guarantees of the thresholding algorithm for the Co-Sparse analysis model, submitted to IEEE Trans. on Information Theory.

[27] R. Giryes, S. Nam, R. Gribonval, M. E. Davies, Iterative cosparse projection algorithms for the recovery of cosparse vectors, in: The 19th European Signal Processing Conference (EUSIPCO-2011), Barcelona, Spain, 2011.

[28] R. Giryes, M. Elad, CoSaMP and SP for the cosparse analysis model, in: The 20th European Signal Processing Conference (EUSIPCO-2012), Bucharest, Romania, 2012.

[29] T. Blumensath, M. Davies, Sampling theorems for signals from the union of finite-dimensional linear subspaces, IEEE Trans. Inf. Theory. 55 (4) (2009) 1872 –1882.

[30] S. Mendelson, A. Pajor, N. Tomczak-Jaegermann, Uniform uncertainty principle for bernoulli and subgaussian ensembles, Constructive Approximation 28 (2008) 277–289.

[31] F. Krahmer, R. Ward, New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property, SIAM J. Math. Analysis 43 (3) (2011) 1269–1281.

[32] R. Baraniuk, M. Davenport, R. DeVore, M. Wakin, A simple proof of the restricted isometry property for random matrices, Constructive Approximation 28 (3) (2008) 253–263.

[33] R. Gribonval, M. E. Pfetsch, A. M. Tillmann, Projection onto the k-cosparse set is NP-hard, Unpuhlished draft, 2012.

[34] T. Han, S. Kay, T. Huang, Optimal segmentation of signals and its application to image denoising and boundary feature extraction, in: International Conference on Image Processing, 2004. ICIP '04., Vol. 4, 2004, pp. 2693 – 2696 Vol. 4.

[35] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused Lasso, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (1) (2005) 91–108.

[36] A. Kyrillidis, V. Cevher, Recipes on hard thresholding methods, in: Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2011 4th IEEE International Workshop on, 2011, pp. 353 –356.

[37] T. Blumensath, Accelerated iterative hard thresholding, Signal Processing 92 (3) (2012) 752 – 756.

[38] M. Rudelson, R. Vershynin, Non-asymptotic theory of random matrices: extreme singular values, in: International Congress of Mathematicans, 2010.

[39] E. J. Candès, The restricted isometry property and its implications for compressed sensing, Comptes-rendus de l'Académie des Sciences, Paris, Series I 346 (910) (2008) 589 – 592.
[40] D. Needell, R. Ward, Stable image reconstruction using total variation minimization, to appear in SIAM J. Imaging Sciences.
[41] D. L. Donoho, J. Tanner, Counting faces of randomly-projected polytopes when the projection radically lowers dimension, J. of the AMS (2009) 1–53.